ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΥΠΟΥΡΓΕΙΟ ΔΙΟΙΚΗΤΙΚΗΣ ΜΕΤΑΡΡΥΘΜΙΣΗΣ & ΗΛΕΚΤΡΟΝΙΚΗΣ ΔΙΑΚΥΒΕΡΝΗΣΗΣ

ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ & ΑΥΤΟΔΙΟΙΚΗΣΗΣ

ΙΝΣΤΙΤΟΥΤΟ ΕΠΙΜΟΡΦΩΣΗΣ

ΕΣΠΑ 2007-2013

Ε.Π. ΔΙΟΙΚΗΤΙΚΗ ΜΕΤΑΡΡΥΘΜΙΣΗ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΠΙΜΟΡΦΩΤΙΚΟ ΠΡΟΓΡΑΜΜΑ

«ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΜΕ ΣΤΑΤΙΣΤΙΚΑ ΠΑΚΕΤΑ»

ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ – ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ 3

2 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ 16

3 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ 18

4 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΤΑ ΕΝΑ ΠΑΡΑΓΟΝΤΑ 29

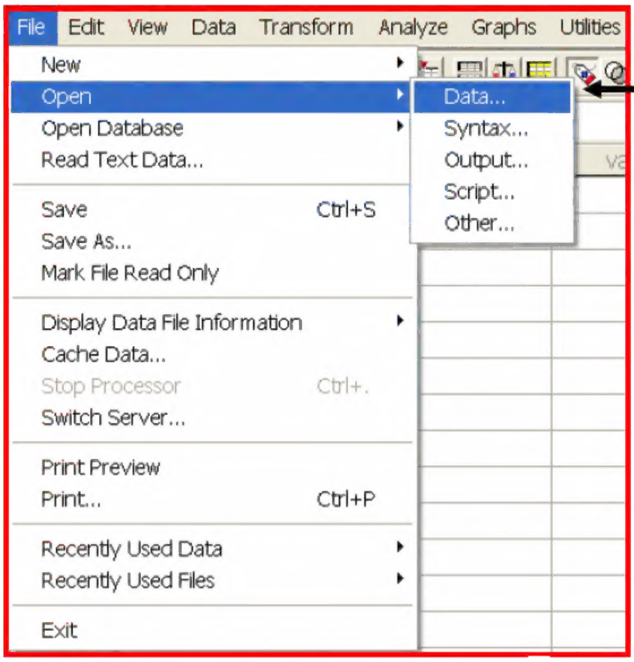
5 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΕΛΕΓΧΟΣ ΑΝΕΞΑΡΤΗΣΙΑΣ 35

6 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΣΥΣΧΕΤΙΣΗ 41

7 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ 46

1 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ – ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ

α) Εισαγωγή Δεδομένων στο SPSS εάν υπάρχουν σε κάποιο άλλο αρχείο.

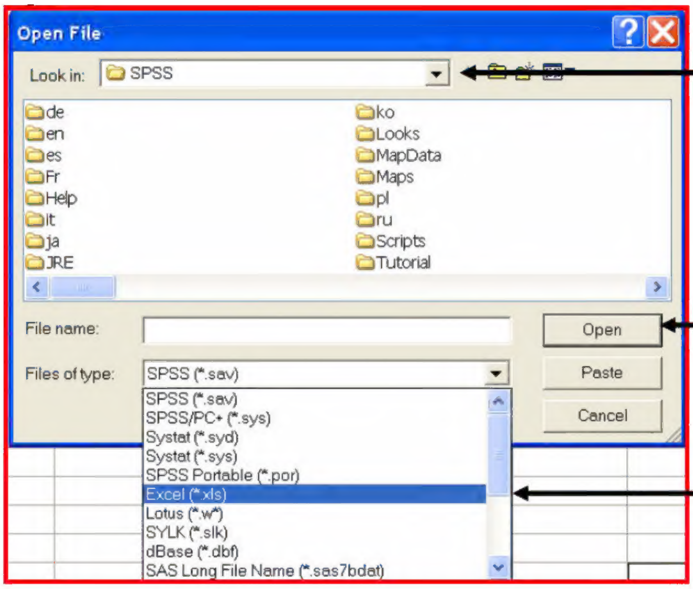


<u>Επιλέγουμε</u>:

File....

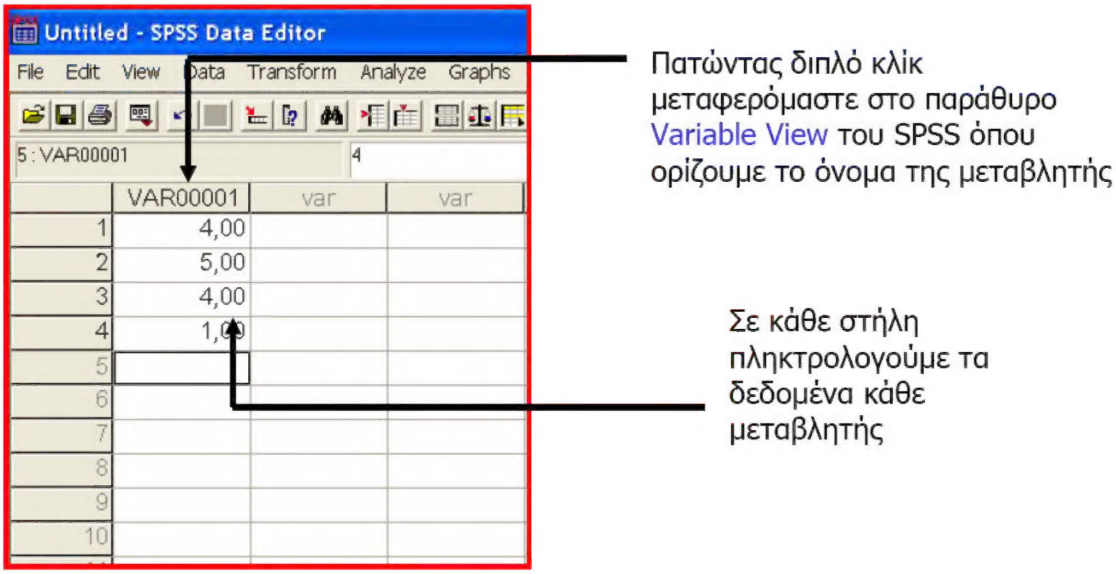
Open....

Data........



Αναζητούμε το αρχείο, επιλέγουμε από τη λίστα τον τύπο αρχείου που είναι 'σωσμένα' τα δεδομένα μας (π.χ. Excel file) και πατάμε **Open.....**

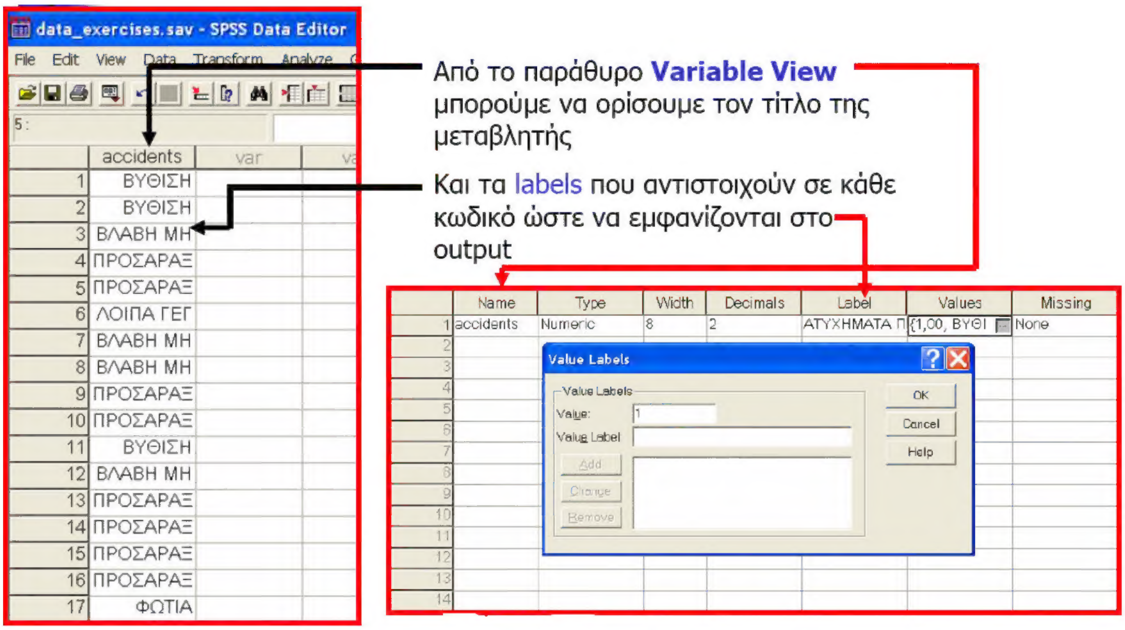
β) Πληκτρολόγηση των δεδομένων στο Data View



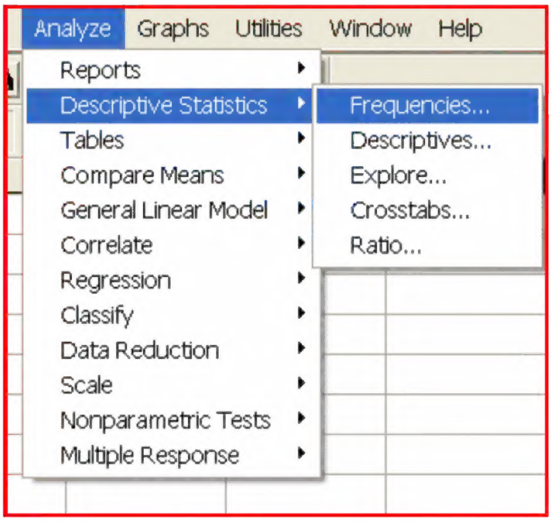
Στη συνέχεια, αναλύουμε τα δεδομένα του παρακάτω πίνακα, όπου δίνονται τα ναυτικά ατυχήματα Ελληνικών εμπορικών πλοίων για το έτος 1993 ανά κατηγορία γεγονότος που προκάλεσε το ατύχημα. Οι κωδικοί αντιστοιχούν σε 1:Βύθιση, 2: Προσάραξη, 3:Φωτιά, 4: Βλάβη μηχανής, 5: Σύγκρουση, 6: Λοιπά γεγονότα. Ο συνολικός αριθμός ατυχημάτων είναι 42 (Πηγή: Στατιστική εμπορικής ναυτιλίας, ΕΣΥΕ).

1	1	4	2	2	6
4	4	2	2	1	4
2	2	2	2	3	4
4	4	4	4	4	4
4	4	5	2	2	2
2	2	4	4	3	2
4	4	4	4	6	2

Εισάγουμε τα δεδομένα



Κατασκευάζουμε τον αντίστοιχο πίνακα συχνοτήτων



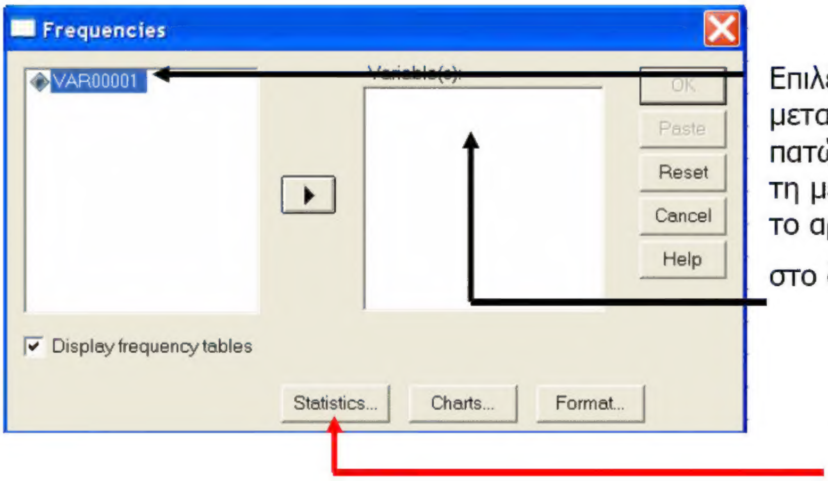
Επιλέγουμε:

Analyze....

Descriptive Statistics...

Frequencies...

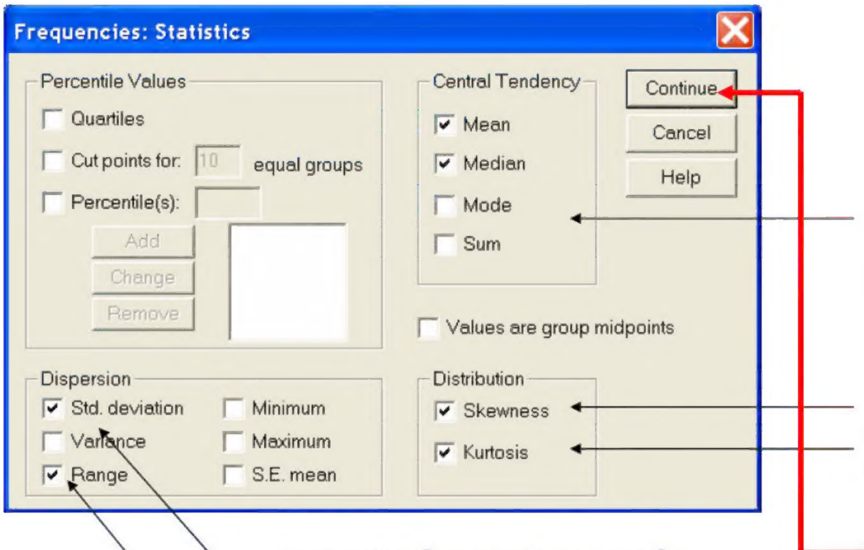
Το πλαίσιο διαλόγου που εμφανίζεται είναι:



Επιλέγουμε τη μεταβλητή και πατώντας το βέλος τη μεταφέρουμε από το αριστερό πλαίσιο στο δεξιό πλαίσιο

Επιλέγουμε:

Statistics.......



Επιλογή **μέτρων θέσης** (π.χ. Μέσος, διάμεσος)

Επιλογή μέτρων **ασυμμετρίας** και **κύρτωσης**

Επιλογή **μέτρων διασποράς** (π.χ. Τυπική απόκλιση, εύρος)

Επιλέγουμε:

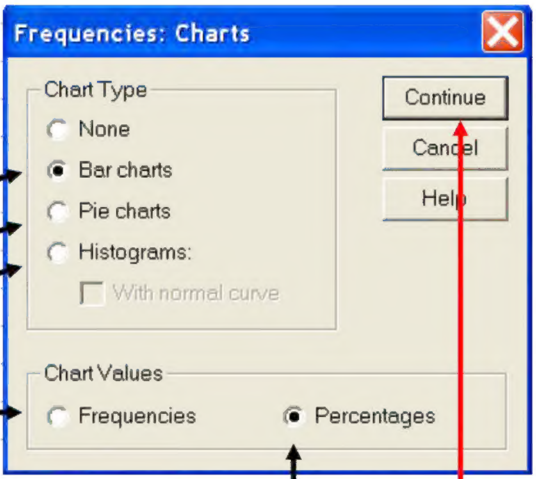
Continue....

Μπορούμε παράλληλα με τον πίνακα συχνοτήτων να απεικονίσουμε γραφικά τα δεδομένα μας επιλέγοντας:

Ραβδόγραμμα

Διάγραμμα πίτας

Ιστόγραμμα

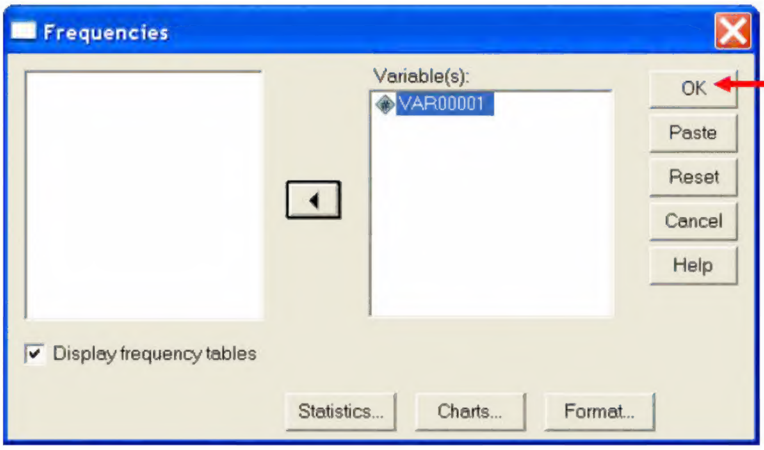


Επιλέγουμε εάν θέλουμε τα αποτελέσματα στα γραφήματα να εμφανίζονται ως **συχνότητες** ή **ποσοστά**

<u>Επιλέγουμε</u>

Continue....

<u>Επιστρέφουμε ξανά στο βασικό παράθυρο</u>:



Επιλέγουμε:

OK....

ΑΤΥΧΗΜΑΤΑ ΠΛΟΙΩΝ ΣΤΗ ΘΑΛΑΣΣΑ

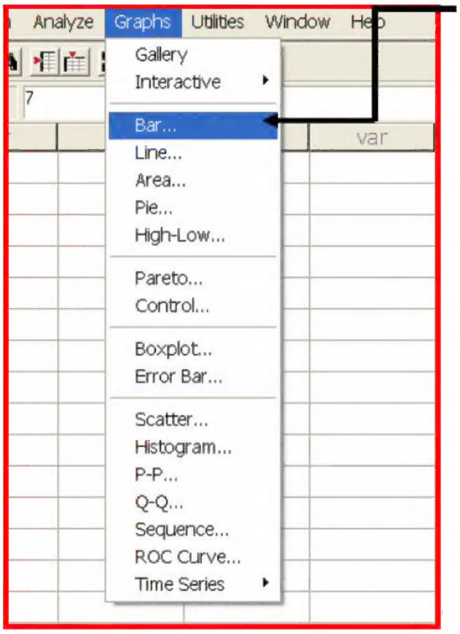
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΒΥΘΙΣΗ	3	7,1	7,1	7,1
	ΠΡΟΣΑΡΑΞΗ	15	35,7	35,7	42,9
	ΦΩΤΙΑ	2	4,8	4,8	47,6
	ΒΛΑΒΗ ΜΗΧΑΝΗΣ	19	45,2	45,2	92,9
	ΣΥΓΚΡΟΥΣΗ	1	2,4	2,4	95,2
	ΛΟΙΠΑ ΓΕΓΟΝΟΤΑ	2	4,8	4,8	100,0
	Total	42	100,0	100,0	

Συχνότητα

Σχετική συχνότητα

Αθροιστική Σχετική Συχνότητα

Κατασκευάζουμε το αντίστοιχο Ραβδόγραμμα.

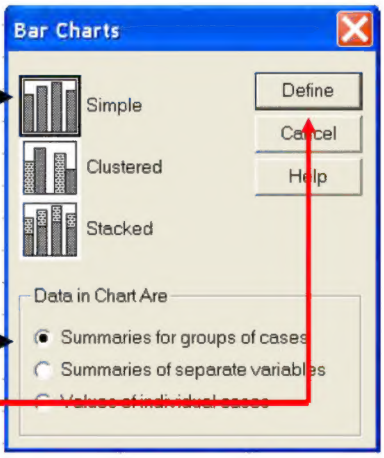


Επιλέγουμε:

Graphs....

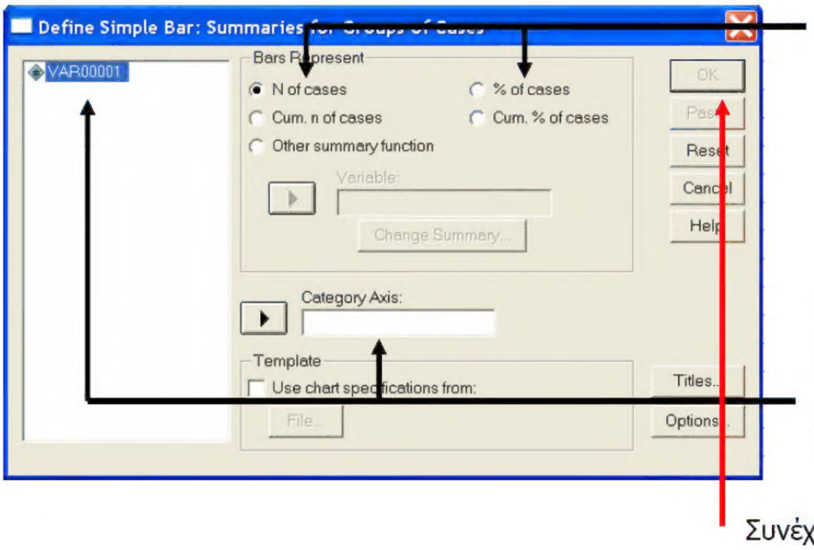
Bar....

Στο επόμενο παράθυρο επιλέγουμε:



Συνεχίζουμε επιλέγοντας:

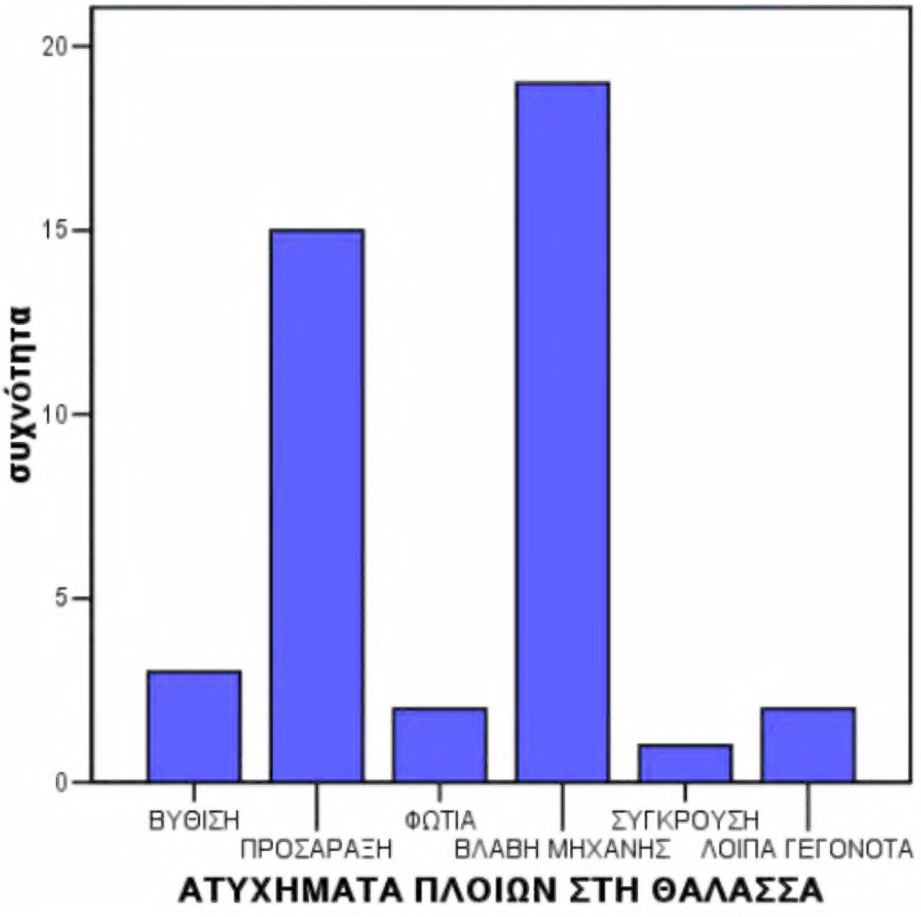
Μεταφερόμαστε στο επόμενο παράθυρο:



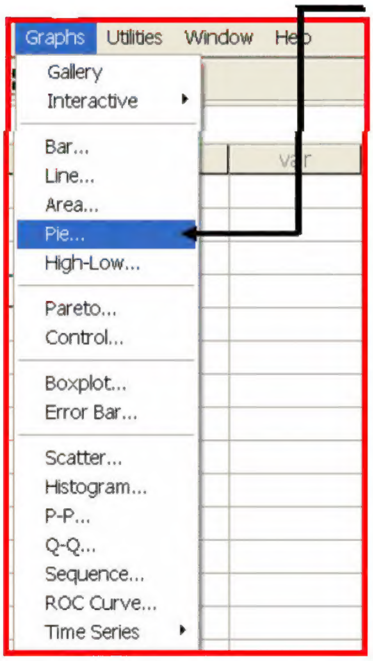
Επιλέγουμε εάν θέλουμε το ύψος της μπάρας να αντιπροσωπεύει **ποσοστό** ή **συχνότητα**

Μεταφέρουμε τη μεταβλητή επιλέγοντάς την και πατώντας το βέλος στο πλαίσιο **Category Axes**

Συνέχεια πατώντας: **OK**



Κατασκευάζουμε διάγραμμα πίτας.



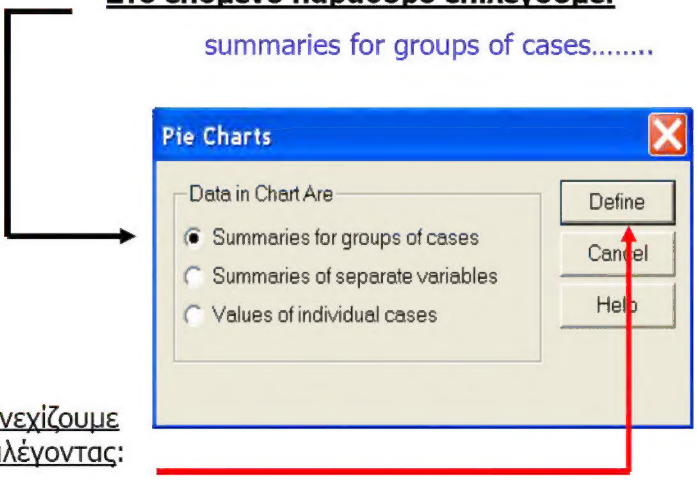
Επιλέγουμε:

Graphs....

Pie....

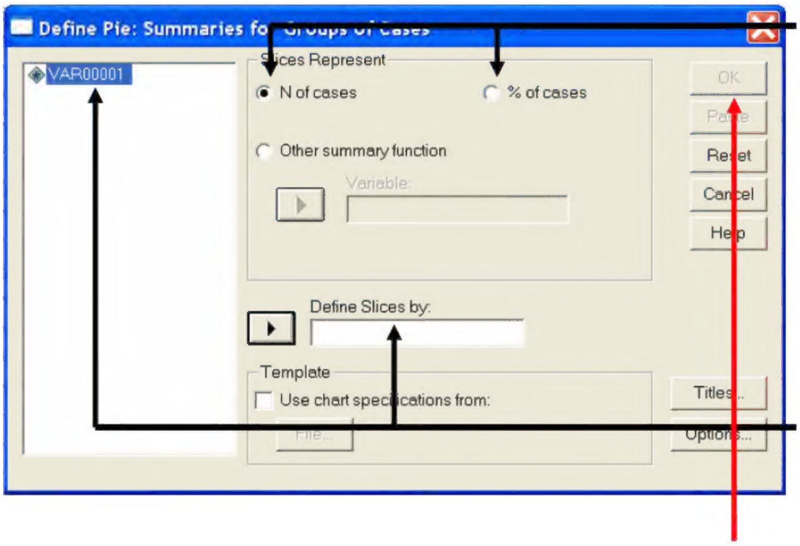
Στο επόμενο παράθυρο επιλέγουμε:

summaries for groups of cases........



Συνεχίζουμε επιλέγοντας:

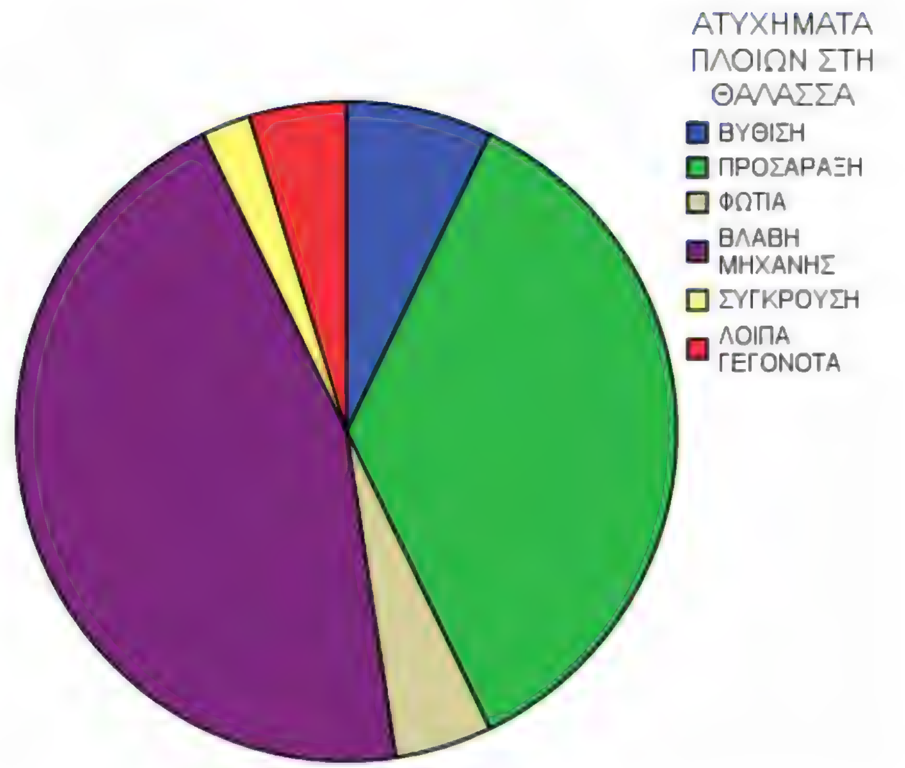
Εμφανίζεται το ακόλουθο παράθυρο διαλόγου.



Επιλέγουμε εάν θέλουμε κάθε τμήμα του κυκλικού διαγράμματος να αντιπροσωπεύει **ποσοστό** ή **συχνότητα**

Μεταφέρουμε τη μεταβλητή επιλέγοντάς την και πατώντας το βέλος στο πλαίσιο: **Define Slices by....**

Συνέχεια πατώντας: **OK**

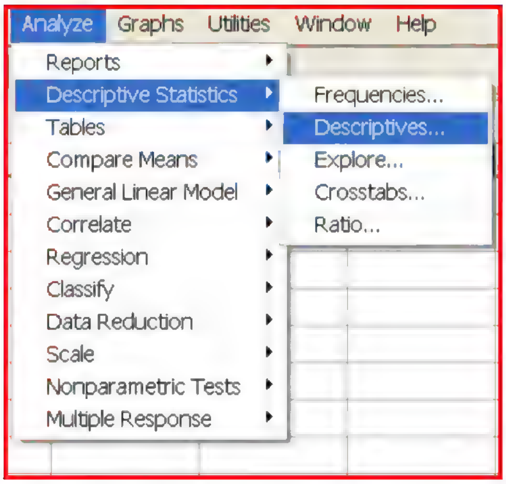


Αντίστοιχα τα βήματα που ακολουθούμε για να αναλύσουμε μία συνεχή μεταβλητή είναι τα ακόλουθα.

Εισάγουμε τα δεδομένα ή ανοίγουμε ένα υπάρχον αρχείο από το κατάλογο του SPSS.

Από το τη βασική ράβδο και τη επιλογή Graph, μπορούμε να κατασκευάζουμε τα αντίστοιχα διαγράμματα (Boxplot – Histogram).

Στη συνέχεια μπορούμε να υπολογίσουμε τα μέτρα θέσης – διασποράς που μας ενδιαφέρουν.

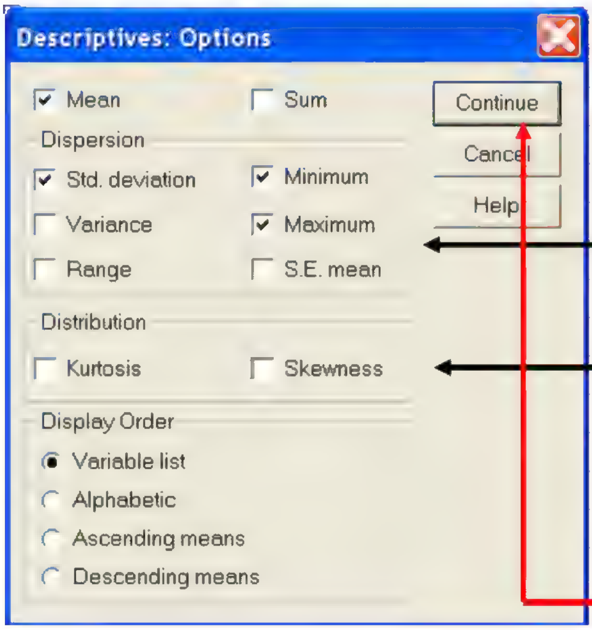


<u>**Επιλέγουμε**</u>:

Analyze....

Descriptive Statistics....

Descriptives...



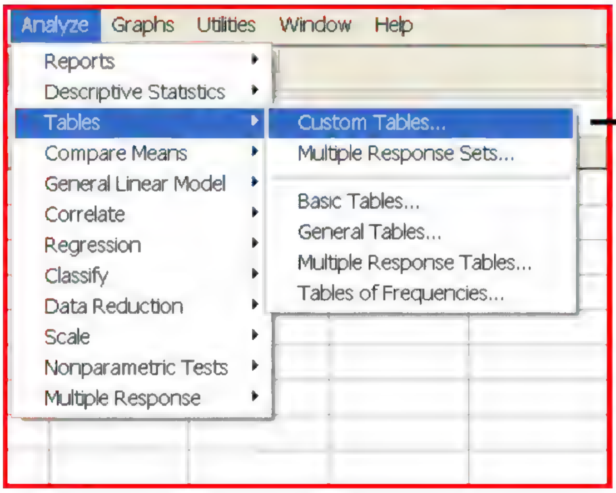
Επιλογή **μέτρων διασποράς**

Επιλέγουμε τα μέτρα **ασυμμετρίας** και **κύρτωσης**

<u>**Επιλέγουμε**</u>:

Continue......

Περισσότερα μέτρα θέσης – διασποράς έχουμε και από την παρακάτω επιλογή.

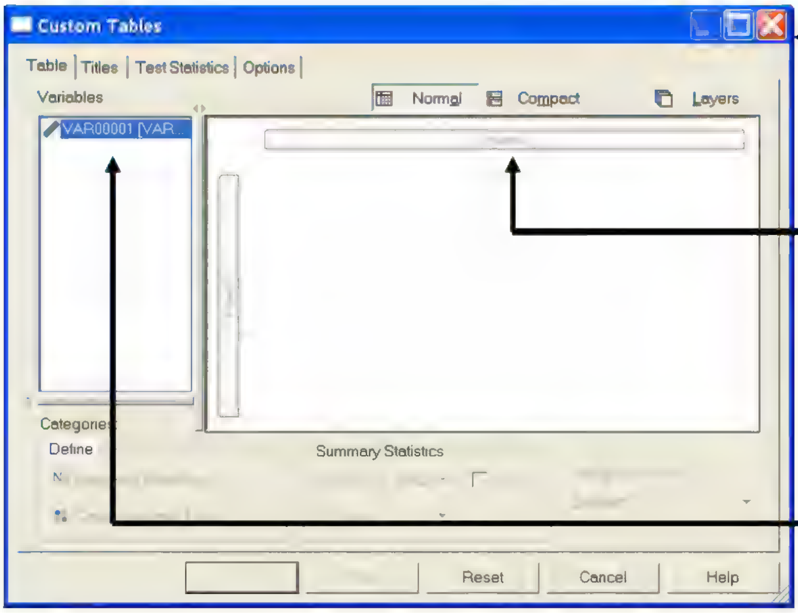


Επιλέγουμε:

Analyze....

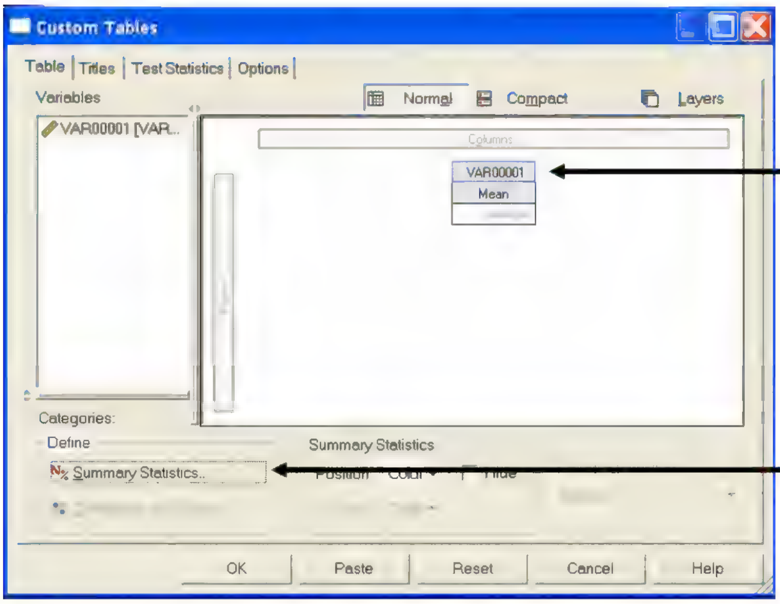
Tables....

Custom Tables....



Θα ανοίξει το διπλανό παράθυρο.

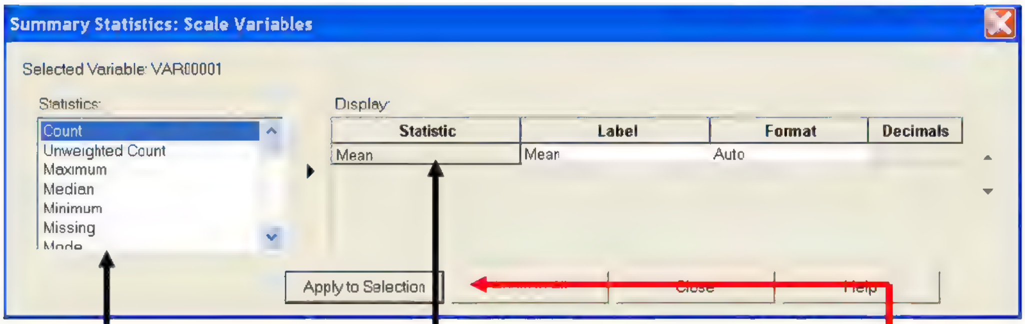
Με Drug-and-Drop μεταφέρουμε τη μεταβλητή από το αριστερό πλαίσιο στο δεξιό πλαίσιο στη θέση **columns**



Η μεταβλητή έχει μεταφερθεί κάτω από τον τίτλο columns

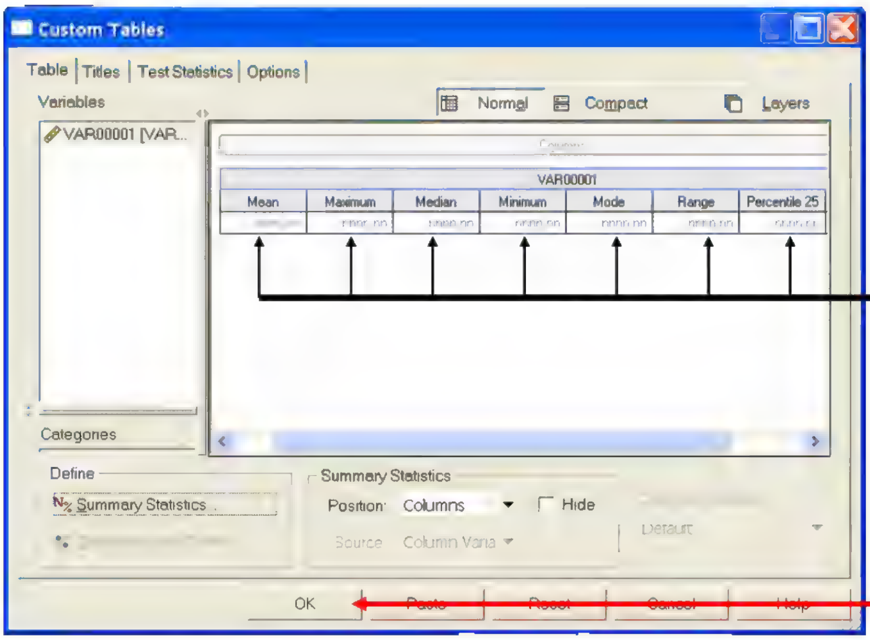
Επιλέγουμε:

Summary statistics.....



Μεταφέρουμε με drug-and-drop ή με το βέλος τα περιγραφικά στατιστικά που θέλουμε από την αριστερή λίστα στο δεξί τμήμα

Τέλος πατάμε apply to selection και επιστρέφουμε στο προηγούμενο παράθυρο



Έχουμε επιλέξει τα ακόλουθα μέτρα θέσης και διασποράς:

Συνεχίζουμε πατώντας **OK....**

2 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Για τα δεδομένα του παραδείγματος να ελεγχθεί η υπόθεση ότι η μέση τιμή είναι μικρότερη από 34:

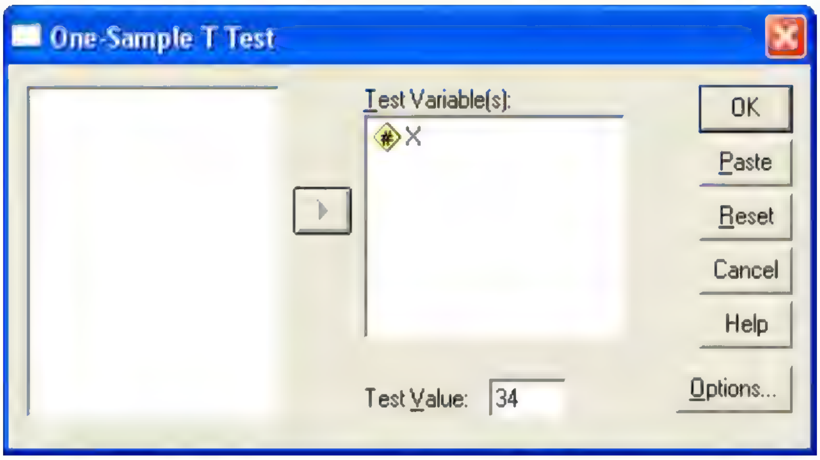
21 51 32 29 42 47 36

Θέλουμε δηλαδή να πραγματοποιήσουμε τον έλεγχο της μορφής:

H_0: μ = 34

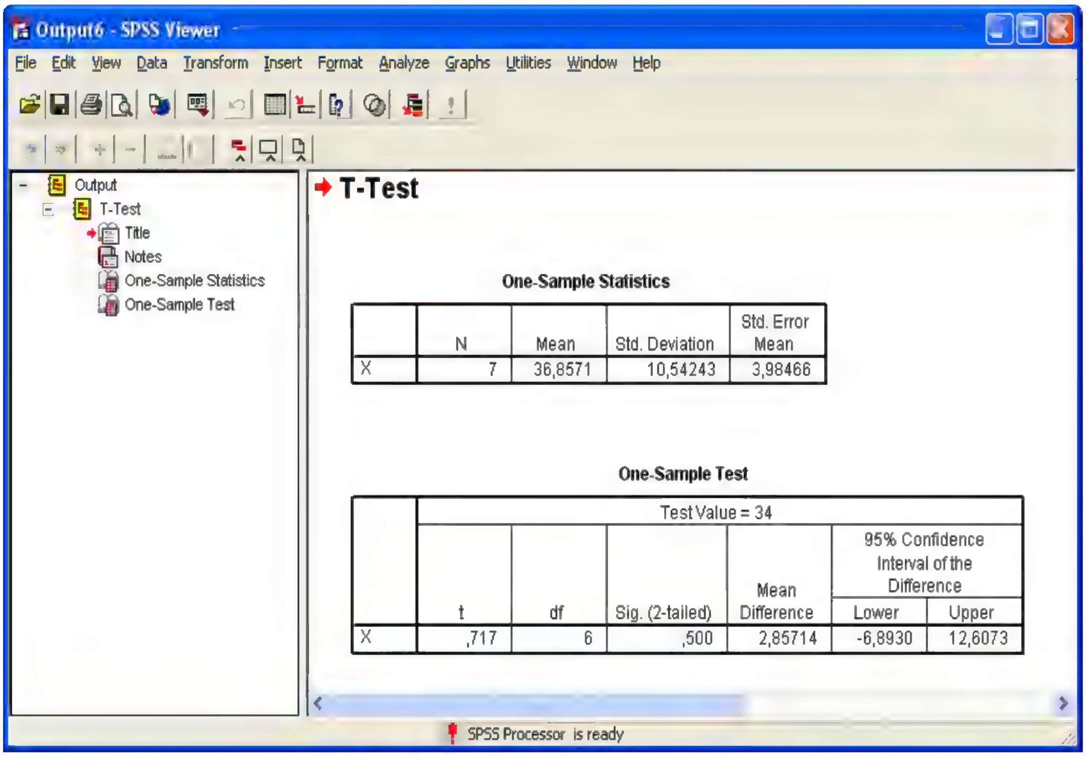
H_1: μ < 34

Αφού καταχωρήσουμε τα δεδομένα ακριβώς όπως στην περίπτωση των διαστημάτων εμπιστοσύνης στο κεντρικό πλαίσιο (καρτέλα ***test value***) ορίζουμε τη τιμή που θέλουμε να γίνει ο έλεγχος.



Εικόνα 1: Επιλογή τιμής για έλεγχο

Το αποτέλεσμα που παίρνουμε φαίνεται στην παρακάτω εικόνα.



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
X	7	36,8571	10,54243	3,98466

One-Sample Test

	Test Value = 34					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	,717	6	,500	2,85714	-6,8930	12,6073

Εικόνα 2: Αποτέλεσμα ελέγχου

Καταρχήν βλέπουμε ότι η μέση τιμή για το δείγμα μας ήταν 36.8, τιμή όχι πολύ μακριά από αυτήν που θέλαμε να ελέγξουμε. Το p-value δίνει τιμή 0.5 για τον αμφίπλευρο έλεγχο. Εμείς έχουμε μονόπλευρο και επομένως το p-value έχει τιμή 0.5/2=0.25, τιμή μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση. Άρα δε μπορούμε να θεωρήσουμε ότι η μέση τιμή είναι μικρότερη από 34.

3 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

Τα στοιχεία που ακολουθούν αφορούν τη μηνιαία κατανάλωση οικογενειών σε γάλα σε τρεις διαφορετικές πόλεις. Από την κάθε πόλη επιλέχτηκε τυχαίου δείγμα τυχαίου μεγέθους. Σκοπός μας είναι να ελέγξουμε κατά πόσο η μηνιαία κατανάλωση γάλακτος είναι ίση μεταξύ των τριών αυτών πόλεων. Έχουμε λοιπόν:

	ΠΟΛΕΙΣ		
#Οικογενειών	**Α**	**Β**	**Γ**
1	24	21	21
2	25	20	22
3	24	21	22
4	24	22	23
5	23		22
6	24		

Πίνακας 1: Μηνιαία κατανάλωση γάλακτος

Σύμφωνα με τα βήματα που έχουμε δει συνοψίζουμε τον παραπάνω πίνακα στη μορφή:

	ΠΟΛΕΙΣ			
#Οικογενειών	**Α**	**Β**	**Γ**	
1	24	21	21	
2	25	20	22	
3	24	21	22	
4	24	22	23	
5	23		22	
6	24			
ΣΥΝΟΛΑ ($y_{i.}$)	144	84	110	338
ΜΕΣΟΙ ($\bar{y}_{i.}$)	24	21	22	
$\sum_i \sum_{j=1}^{n_i} y_{ij}^2$	3458	1766	2422	7646

Πίνακας 2: Πίνακας δεδομένων και μεταβλητών α' τρόπου

Επομένως, για τα αθροίσματα τετραγώνων των αποκλίσεων καταλήγουμε στα εξής:

$$SSB = \sum_{i=1}^{k} \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n} = \left(\frac{144^2}{6} + \frac{84^2}{4} + \frac{110^2}{5} \right) - \frac{(144 + 84 + 110)^2}{15} = 7640 - 7616.27 = 23.73$$

$$SSW = \sum_i \sum_j Y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i} = (3458 + 1766 + 2422) - \left(\frac{144^2}{6} + \frac{84^2}{4} + \frac{110^2}{5} \right) =$$

$$= 7646 - (3456 + 1764 + 2420) = 6$$

$$SST = SSB + SSW = 23.73 + 6 = 29.73$$

Εναλλακτικά, μπορούμε να επιλύσουμε το παραπάνω πρόβλημα με τον ακόλουθο τρόπο:

	ΠΟΛΕΙΣ					
#Οικογενειών	**Α**	$(\mathbf{y}-\overline{\mathbf{y}}_1)^2$	**Β**	$(\mathbf{y}-\overline{\mathbf{y}}_2)^2$	**Γ**	$(\mathbf{y}-\overline{\mathbf{y}}_3)^2$
1	24	0	21	0	21	1
2	25	1	20	1	22	0
3	24	0	21	0	22	0
4	24	0	22	1	23	1
5	23	1			22	0
6	24	0				
ΣΥΝΟΛΑ ($y_{i.}$)	144	2	84	2	110	2
ΜΕΣΟΙ ($\bar{y}_{i.}$)	24		21		22	

Πίνακας 3: Πίνακας δεδομένων και μεταβλητών β' τρόπου

Επίσης, $\bar{y}_{..} = \frac{y_{..}}{n} = \frac{144+84+110}{15} = 22.53$

Για τα αθροίσματα των τετραγώνων των αποκλίσεων παίρνουμε τις εξής τιμές:

$$SSW = \sum_i \sum_j \left(y_{ij} - \bar{y}_{i.} \right)^2 = 2+2+2 = 6$$

$$SSB = \sum_i \sum_j \left(\bar{y}_{i.} - \bar{y}_{..} \right)^2 = \sum_i n_i \left(\bar{y}_{i.} - \bar{y}_{..} \right)^2 =$$

$$= 6 \cdot (24-22.53)^2 + 4 \cdot (21-22.53)^2 + 5 \cdot (22-22.53)^2 = 23.73$$

$$SST = SSW + SSB = 29.73$$

Όποιον και από τους δύο παραπάνω τρόπους διαλέξουμε θα καταλήξουμε στην εξής μορφή του πίνακα ανάλυσης διακύμανσης:

Αιτία Διασποράς	**Άθροισμα Τετραγώνων**	**Βαθμοί Ελευθερίας**	**Μέσα Τετραγωνικά Λάθη**	**F (κάτω από την H_0)**
Μεταξύ Δειγμάτων (between samples)	23.73	3-1	23.73/2=11.87	11.87/0.5=23.74
Εντός Δειγμάτων (within samples)	6	15-3	6/12=0.5	
Σύνολο	29.73	14		

Πίνακας 4: Έλεγχος Ανάλυσης Διακύμανσης παραδείγματος

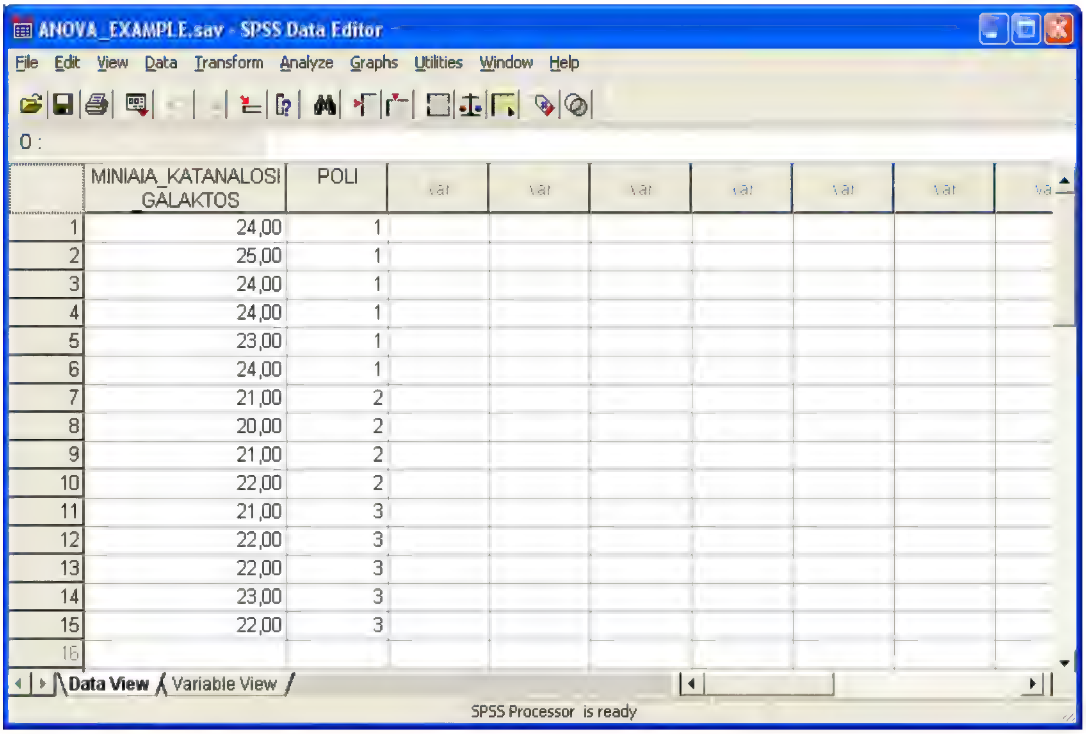
Όπου για α=0.05 έχουμε $F_{(k-1),(n-k),\alpha}= F_{2,12,0.95}=3.88$ από τους πίνακες κατανομής F.

Θέλαμε να ελέγξουμε την υπόθεση:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{τουλάχιστον δύο μέσοι διαφέρουν.}$$

Αφού F > $F_{2,12,0.95}$ απορρίπτουμε τη μηδενική υπόθεση περί ισότητας των τριών μέσων. Επομένως θεωρούμε ότι τουλάχιστον μια μηνιαία κατανάλωση γάλακτος αναφορικά με τις τρεις πόλεις διαφέρει από τις άλλες δύο.

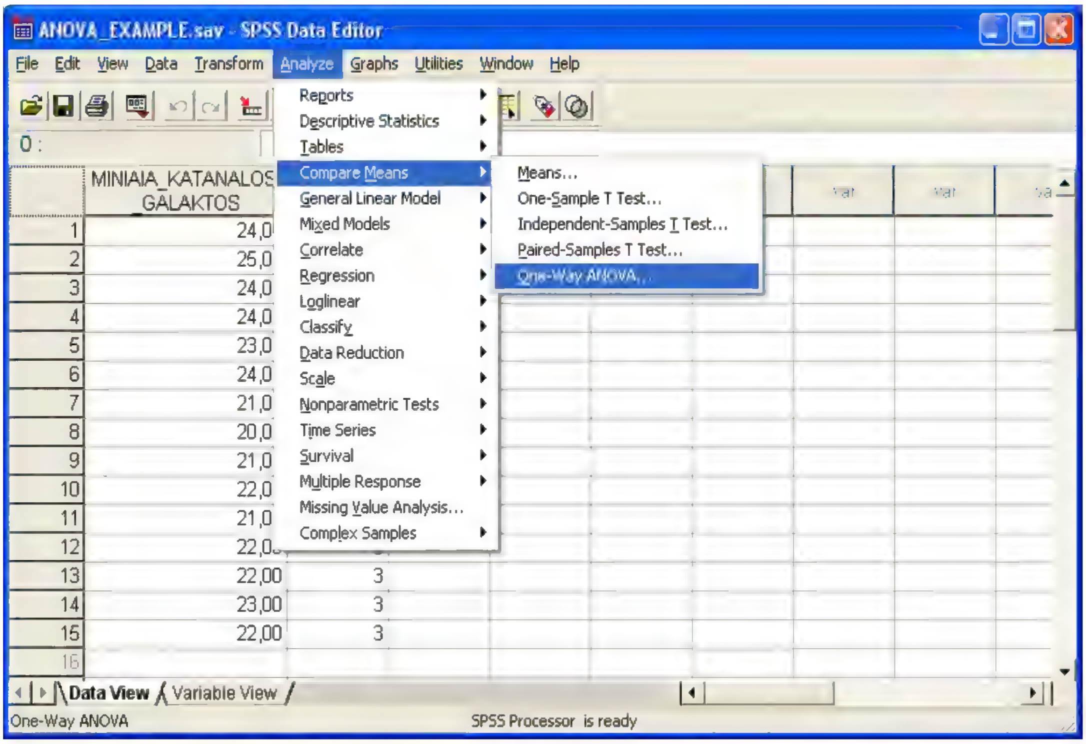
Εφαρμόζουμε, τα παραπάνω με τη χρήση του SPSS. Εισάγουμε όλα τα δεδομένα στην πρώτη στήλη και την ονομάζουμε π.χ. ***MINIAIA_KATANALOSI_GALAKTOS***. Αυτή αποτελεί την εξαρτημένη μας μεταβλητή. Στην διπλανή στήλη θα πρέπει να εισάγουμε τις αντίστοιχες πόλεις από τις οποίες πήραμε το δείγμα. Επειδή, ωστόσο, η μέθοδος αυτή χρειάζεται αριθμητικά δεδομένα σε όλες τις τιμές που αφορούν την πόλη Α αντιστοιχούμε την τιμή 1. Ομοίως στις τιμές των δεδομένων της πόλης Β αντιστοιχούμε τον αριθμό 2 και σε αυτές της πόλης Γ τον αριθμό 3. Αυτό θα συνεχιζόταν με τον ίδιο τρόπο αν είχαμε και άλλα δείγματα (sample). Τη νέα αυτή μεταβλητή την ονομάζουμε ***POLI***. Επομένως, έχουμε τα δεδομένα στην μορφή:



	MINIAIA_KATANALOSI GALAKTOS	POLI
1	24,00	1
2	25,00	1
3	24,00	1
4	24,00	1
5	23,00	1
6	24,00	1
7	21,00	2
8	20,00	2
9	21,00	2
10	22,00	2
11	21,00	3
12	22,00	3
13	22,00	3
14	23,00	3
15	22,00	3

Εικόνα 1 : Εισαγωγή δεδομένων στο SPSS

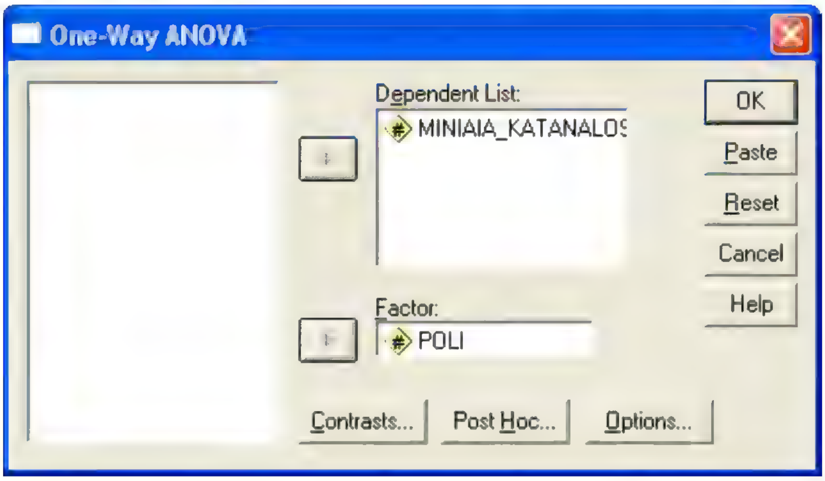
Το επόμενο βήμα είναι να εφαρμόσουμε τη μέθοδο. Ακολουθούμε αναλυτικά τα βήματα:



Εικόνα 2 : Βήματα εφαρμογής ANOVA

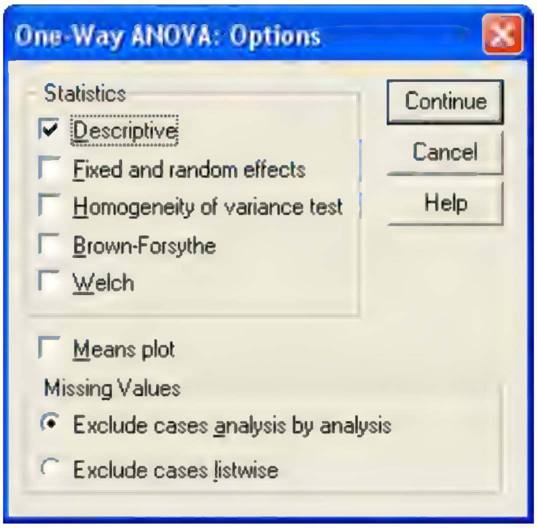
Στο κεντρικό menu της μεθόδου θα πρέπει να καθορίσουμε ποια μεταβλητή είναι η εξαρτημένη και ως προς ποια μεταβλητή θέλουμε να ελέγξουμε τον έλεγχο περί ισότητας των μέσων (παράγοντας).

Στο παράδειγμα μας θέλουμε να ελέγξουμε την υπόθεση περί ισότητας των μέσων της μηνιαίας κατανάλωσης γάλακτος (εξαρτημένη μεταβλητή) ως προς τις τρεις διαφορετικές πόλεις (παράγοντας). Επομένως έχουμε:



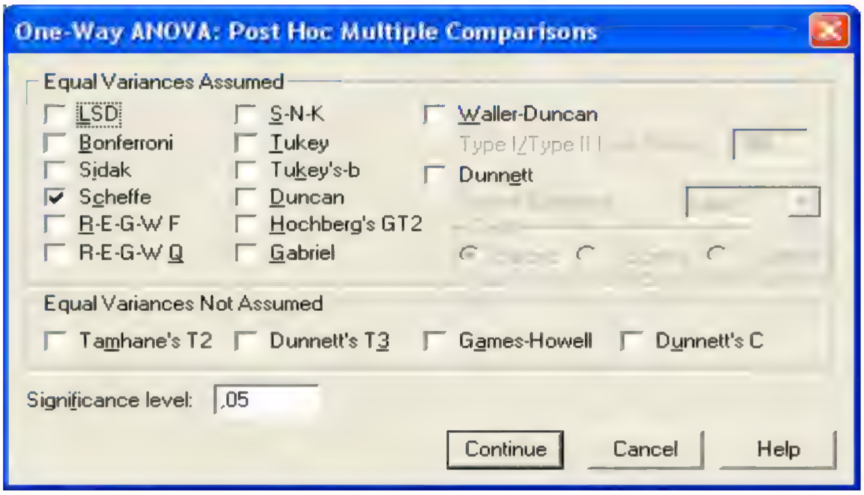
Εικόνα 3 : Επιλογή εξαρτημένης επιλογής και παράγοντα

Στην καρτέλα ***Options*** μπορούμε να επιλέξουμε κάποια περιγραφικά μέτρα (Descriptive). Αναλυτικά:



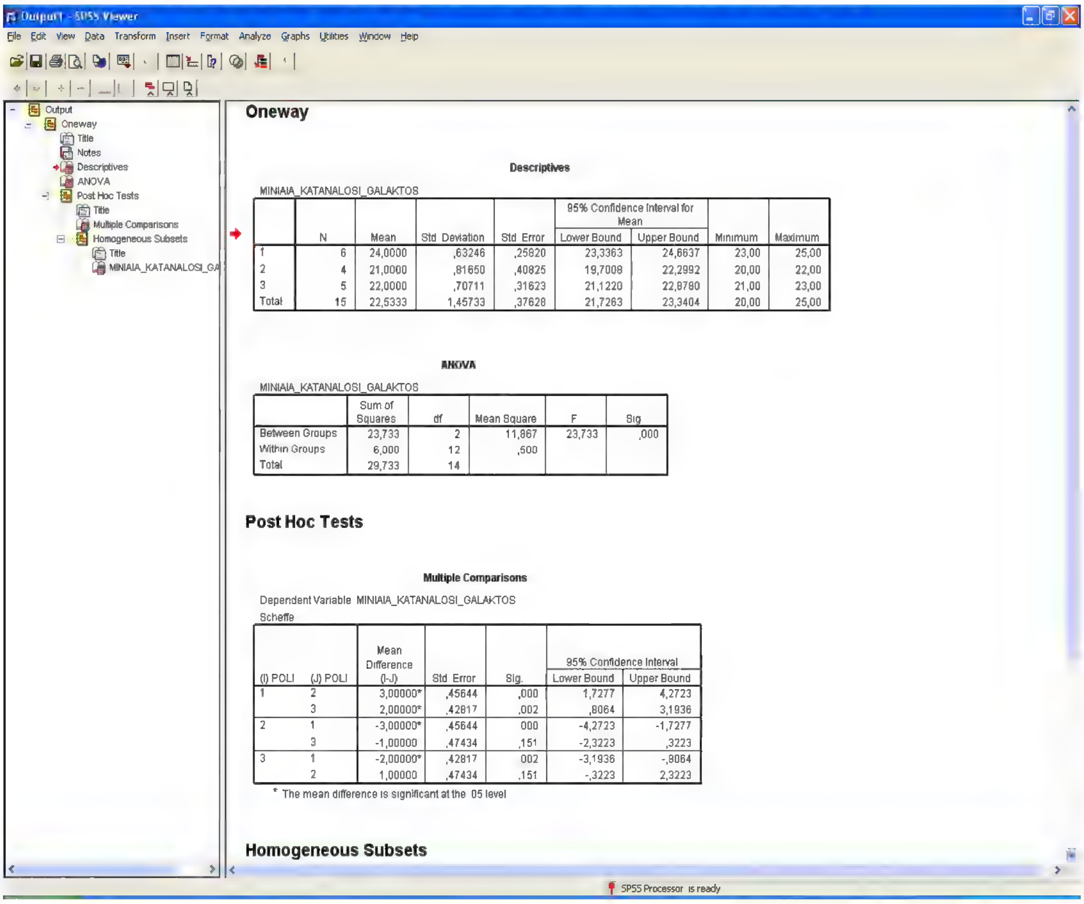
Εικόνα 4 : Καρτέλα ***Options***

Επίσης, στην καρτέλα ***Post Hoc*** έχουμε τη δυνατότητα να επιλέξουμε κάποιο test που εντοπίζει ανάμεσα σε ποιους μέσους υπάρχει διαφορά αν και εφόσον απορριφθεί η μηδενική υπόθεση. Έστω ότι επιλέγουμε τον έλεγχο του Scheffe.



Εικόνα 5 : Καρτέλα ***Post Hoc***

Εκτελώντας λοιπόν την εντολή παίρνουμε τα παρακάτω αποτελέσματα:



Oneway

Descriptives

MINIAIA_KATANALOSI_GALAKTOS

	N	Mean	Std Deviation	Std Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	6	24,0000	,63246	,25820	23,3363	24,6637	23,00	25,00
2	4	21,0000	,81650	,40825	19,7008	22,2992	20,00	22,00
3	5	22,0000	,70711	,31623	21,1220	22,8780	21,00	23,00
Total	15	22,5333	1,45733	,37628	21,7263	23,3404	20,00	25,00

ANOVA

MINIAIA_KATANALOSI_GALAKTOS

	Sum of Squares	df	Mean Square	F	Sig
Between Groups	23,733	2	11,867	23,733	,000
Within Groups	6,000	12	,500		
Total	29,733	14			

Post Hoc Tests

Multiple Comparisons

Dependent Variable MINIAIA_KATANALOSI_GALAKTOS
Scheffe

(I) POLI	(J) POLI	Mean Difference (I-J)	Std Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	3,00000*	,45644	,000	1,7277	4,2723
	3	2,00000*	,42817	,002	,8064	3,1936
2	1	-3,00000*	,45644	000	-4,2723	-1,7277
	3	-1,00000	,47434	,151	-2,3223	,3223
3	1	-2,00000*	,42817	002	-3,1936	-,8064
	2	1,00000	,47434	,151	-,3223	2,3223

* The mean difference is significant at the 05 level

Homogeneous Subsets

Εικόνα 6 : One Way ANOVA

Ο πρώτος πίνακας αφορά κάποια περιγραφικά μέτρα.

Descriptives

MINIAIA_KATANALOSI_GALAKTOS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	6	24,0000	,63246	,25820	23,3363	24,6637	23,00	25,00
2	4	21,0000	,81650	,40825	19,7008	22,2992	20,00	22,00
3	5	22,0000	,70711	,31623	21,1220	22,8780	21,00	23,00
Total	15	22,5333	1,45733	,37628	21,7263	23,3404	20,00	25,00

Εικόνα 7 : Περιγραφικά Μέτρα

Βλέπουμε λοιπόν ότι έχουμε τις 3 πόλεις. Φαίνεται το πλήθος του κάθε δείγματος ανά πόλη, η μέση τιμή στη μηνιαία κατανάλωση γάλακτος, η τυπική απόκλισης, ένα 95% διάστημα εμπιστοσύνης, καθώς επίσης και η μικρότερη και η μέγιστη τιμή, πάλι ανά πόλη.

Ο επόμενος πίνακας είναι και ο βασικός και αποτυπώνει τον πίνακα Ανάλυσης Διακύμανσης

ANOVA

MINIAIA_KATANALOSI_GALAKTOS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	23,733	2	11,867	23,733	,000
Within Groups	6,000	12	,500		
Total	29,733	14			

Εικόνα 8 : Πίνακας Ανάλυσης Διακύμανσης

Προφανώς, ο πίνακας που παίρνουμε ταυτίζεται με αυτόν που είχαμε κατασκευάσει παραπάνω. Εκτός της τιμής F (23.733) το SPSS μας δίνει και την τιμή του επίπεδου σημαντικότητας p που

αντιστοιχεί σε αυτήν και είναι σχεδόν 0. Έτσι έχουμε:

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : τουλάχιστον \quad δύο \quad μέσοι \quad διαφέρουν.$

Απορρίπτουμε τη μηδενική υπόθεση H_0 όταν το p έχει τιμή μικρότερη ή ίση με την τιμή του επιπέδου σημαντικότητας α που έχει γίνει η μέθοδος. Το α έχει ορισθεί στο 0.05 και επειδή $p<0.05$ απορρίπτουμε την μηδενική υπόθεση περί ισότητας των μέσων. Επομένως, έχουμε καταλήξει στο συμπέρασμα ότι τουλάχιστον δύο μέσες τιμές διαφέρουν. Το ερώτημα είναι ποιες.

Για αυτό το λόγο χρησιμοποιήσαμε την επιλογή των ***Post Hoc*** ελέγχων. Έχουμε:

Multiple Comparisons

Dependent Variable: MINIAIA_KATANALOSI_GALAKTOS
Scheffe

(I) POLI	(J) POLI	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	3,00000*	,45644	,000	1,7277	4,2723
	3	2,00000*	,42817	,002	,8064	3,1936
2	1	-3,00000*	,45644	,000	-4,2723	-1,7277
	3	-1,00000	,47434	,151	-2,3223	,3223
3	1	-2,00000*	,42817	,002	-3,1936	-,8064
	2	1,00000	,47434	,151	-,3223	2,3223

*. The mean difference is significant at the .05 level.

Εικόνα 9 : Έλεγχος Sceffe για την ανίχνευση διαφορών στους μέσους

Όπως παρατηρούμε ο έλεγχος συγκρίνει τις μέσες τιμές ανά πόλη. Δηλαδή, πρώτα τη μέση τιμή της Α πόλης με αυτήν της Β, έπειτα της Α με της Γ και τέλος της Β με την Γ.

Σε κάθε μία περίπτωση διενεργείται ο έλεγχος:

$H_0 : \mu_i = \mu_j$

$H_1 : \mu_i \neq \mu_j$

- Στην πρώτη περίπτωση (Α με Β) το p είναι μικρότερο του 0.05 οπότε απορρίπτεται η μηδενική υπόθεση. Οι μέσες τιμές της μηνιαίας κατανάλωσης γάλακτος μεταξύ των πόλεων Α και Β διαφέρουν στατιστικά σημαντικά.

- Στη δεύτερη περίπτωση (Α με Γ) το p είναι μικρότερο του 0.05 (p=0.02) οπότε απορρίπτεται και εδώ η μηδενική υπόθεση. Άρα, οι μέσες τιμές της μηνιαίας κατανάλωσης γάλακτος μεταξύ των πόλεων Α και Γ διαφέρουν στατιστικά σημαντικά.
- Στην Τρίτη περίπτωση (Β με Γ) το p είναι μεγαλύτερο του 0.05 (p=0.151) οπότε δεν απορρίπτεται η μηδενική υπόθεση. Άρα, οι μέσες τιμές της μηνιαίας κατανάλωσης γάλακτος μεταξύ των πόλεων Β και Γ δεν διαφέρουν στατιστικά σημαντικά.

Μπορούμε βέβαια αντί για το test του Scheffe να εφαρμόσουμε οποιονδήποτε άλλο από τους ελέγχους που μας παρέχονται.

4 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΤΑ ΕΝΑ ΠΑΡΑΓΟΝΤΑ

Η παρουσία βλαβερών εντόμων στις καλλιέργειες είναι γνωστό ότι προκαλεί συχνά σημαντικές ζημιές στις σοδειές. Μια από τις λύσεις που έχουν προταθεί για την αντιμετώπιση αυτού του προβλήματος είναι η χρησιμοποίηση υλικών τα οποία έχουν στην επιφάνεια τους ειδική κολλώδη ουσία και έχουν την ιδιότητα να παγιδεύουν τα έντομα. Βοτανολόγος ενδιαφέρεται να εξετάσει αν το χρώμα που έχει η κολλώδης επιφάνεια ενδέχεται να επηρεάζει τον αριθμό των εντόμων που κολλάνε σε αυτήν, έτσι ώστε να εντοπίσει ποιο από τα χρώματα οδηγεί στην αποτελεσματικότερη αντιμετώπιση του προβλήματος. Για το λόγο αυτό ετοίμασε 24 κολλώδεις επιφάνειες ιδίων διαστάσεων οι οποίες διέφεραν μόνο ως προς το χρώμα τους: υπήρχαν 6 πράσινες, 6 κόκκινες, 6 κίτρινες και 6 μπλε. Στη συνέχεια επέλεξε μια καλλιεργήσιμη έκταση μεγάλου μεγέθους και σε 24 τυχαία επιλεγμένα σημεία αυτής τοποθέτησε τις 24 επιφάνειες. Η επιφάνεια που τοποθετήθηκε σε κάθε σημείο επιλέχθηκε τυχαία. 24 ώρες μετά, έγινε η καταγραφή των εντόμων που παγιδεύτηκαν σε κάθε επιφάνεια. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα.

	ΧΡΩΜΑ ΕΠΙΦΑΝΕΙΑΣ			
	ΠΡΑΣΙΝΟ	ΚΟΚΚΙΝΟ	ΚΙΤΡΙΝΟ	ΜΠΛΕ
ΑΡΙΘΜΟΣ ΕΝΤΟΜΩΝ	59	24	41	42
	48	41	55	42
	38	43	38	50
	46	34	47	40
	52	44	44	56
	55	31	38	40

Να συγγράψετε αναφορά στην οποία θα αναλύετε τα παραπάνω δεδομένα με την καταλληλότερη κατά την άποψη σας στατιστική μέθοδο και θα δίνετε στο βοτανολόγο όλα τα απαραίτητα στοιχεία που προκύπτουν για την επίδραση του χρώματος πάνω στην προσέλκυση των εντόμων.

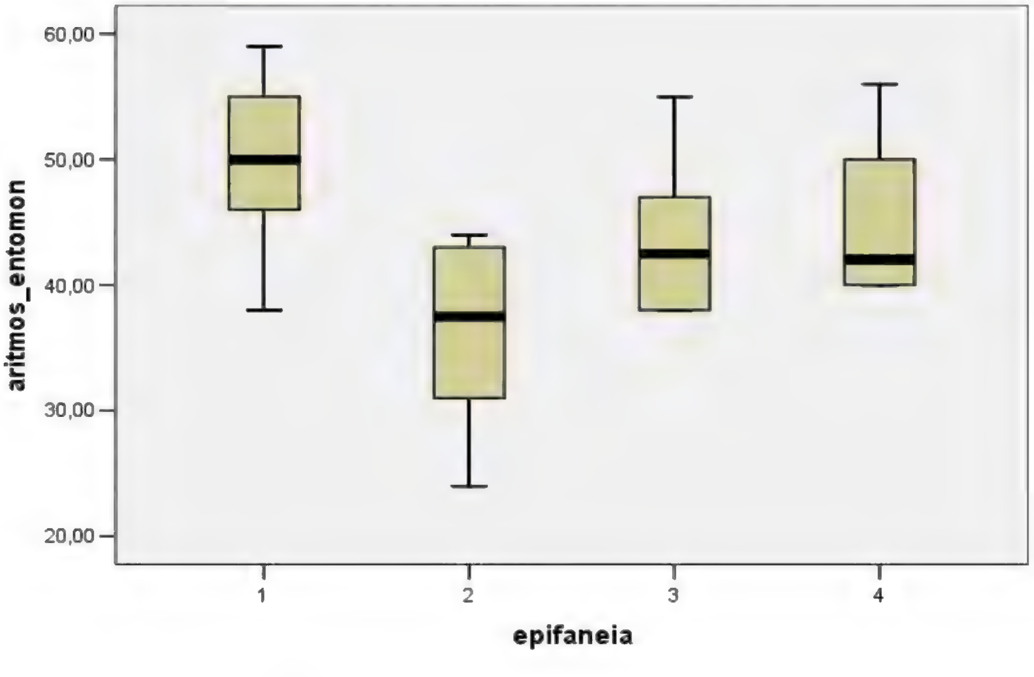
Επίσης, να αναφερθείτε σε εναλλακτικούς σχεδιασμούς του πειράματος που ενδεχομένως θα διαφοροποιούσαν την εγκυρότητα των αποτελεσμάτων.

ΠΡΟΤΕΙΝΟΜΕΝΗ ΑΠΑΝΤΗΣΗ

Από τη φύση του προβλήματος αντιλαμβανόμαστε ότι η καλύτερη μέθοδος αντιμετώπισης του είναι η Ανάλυση Διακύμανσης κατά ένα Παράγοντα (One Way ANOVA). Επί της ουσίας δηλαδή, θέλουμε να εξακριβώσουμε αν το χρώμα παίζει ρόλο στον αριθμό των εντόμων που παγιδεύονται στην κολλώδη επιφάνεια. Υπάρχει δηλαδή κάποιο συγκεκριμένο χρώμα που έλκει περισσότερο τα έντομα αυτά και αν ναι ποιο είναι αυτό;

Το πρώτο που απαιτείται είναι να εισαχθούν τα δεδομένα στο SPSS. Καταχωρούνται σε δύο στήλες, η πρώτη, ***'aritmos_entomon'*** περιλαμβάνει τα δεδομένα και η δεύτερη στήλη, ***'epifaneia'*** είναι η κωδικοποίηση των δεδομένων ανάλογα με το χρώμα. Δηλαδή, το 1 αντιστοιχεί στο πράσινο, το 2 στο κόκκινο, το 3 στο κίτρινο και το 4 στο μπλε.

Δημιουργείται ένα γράφημα για μια πρώτη εικόνα των δεδομένων.



Εικόνα 1: Box plot γράφημα των δεδομένων

Με μια πρώτη ματιά φαίνεται ότι ο διάμεσος αριθμός των εντόμων που παγιδεύονται στα χρώματα πράσινο, κίτρινο και μπλε είναι περίπου ίδιος, σε αντίθεση με το κόκκινο χρώμα, όπου φαίνεται να είναι μικρότερος. Το ερώτημα είναι βέβαια να υπάρχουν στατιστικά σημαντικές διαφορές ώστε να θεωρηθεί ότι κάποιο χρώμα είναι πιο αποτελεσματικό από τα υπόλοιπα.

Η ANOVA βασίζεται στον έλεγχο της υπόθεσης:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

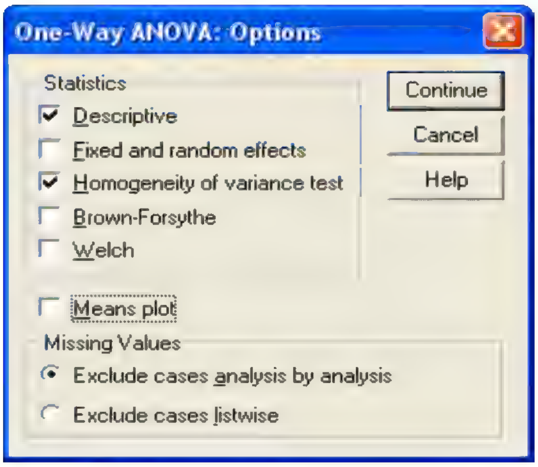
H_1: τουλάχιστον ένα μ_i διαφέρει

Για την εφαρμογή του ελέγχου χρειάζεται να ελεγθούν οι υποθέσεις:

- Ανεξαρτησία
- Ομοσκεδαστικότητα
- Κανονικότητα

Η ανεξαρτησία μπορεί να επιτευχθεί εάν κατά το σχεδιασμό του πειράματος έχει προβλεφθεί να υπάρχει τυχαιοποίηση κάτι που συνεπάγεται και ανεξαρτησία των παρατηρήσεων.

Για τον έλεγχο της ομοσκεδαστικότητας από την καρτέλα ***'Options'*** επιλέγουμε:



Εικόνα 2: Έλεγχος ομοσκεδαστικότητας

Το αποτέλεσμα δίνει:

Test of Homogeneity of Variances

aritmos_entomon

Levene Statistic	df1	df2	Sig.
,234	3	20	,872

Εικόνα 3: Αποτέλεσμα ελέγχου ομοσκεδαστικότητας

Αφού το p-value (p-value=0,872) έχει τιμή μεγαλύτερη από 0.05 δεν απορρίπτεται η μηδενική υπόθεση και επομένως ισχύει και αυτή η υπόθεση.

Τέλος, για την κανονικότητα μπορεί να εφαρμοστεί ο έλεγχος Kolmogorov –Smirnov.

One-Sample Kolmogorov-Smirnov Test

		aritmos_ entomon
N		24
Normal Parameters[a,b]	Mean	43,6667
	Std. Deviation	8,27078
Most Extreme Differences	Absolute	,122
	Positive	,109
	Negative	-,122
Kolmogorov-Smirnov Z		,596
Asymp. Sig. (2-tailed)		,870
Exact Sig. (2-tailed)		,828
Point Probability		,000

a. Test distribution is Normal.

b. Calculated from data.

Εικόνα 4: Αποτέλεσμα ελέγχου Kolmogorov - Smirnov

Και εδώ φαίνεται να ισχύει η υπόθεση, αφού το p-vale είναι μεγαλύτερο από 0.05

Μπορεί να προχωρήσει η εφαρμογή του ελέγχου. Δίνει τα ακόλουθα αποτελέσματα:

ANOVA

aritmos_entomon

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	564,333	3	188,111	3,729	,028
Within Groups	1009,000	20	50,450		
Total	1573,333	23			

Εικόνα 5: One Way ANOVA

Παρατηρούμε ότι το P-value έχει τιμή 0.028 και επομένως σε επίπεδο σημαντικότητα α=0.05 απορρίπτεται η μηδενική υπόθεση ότι όλοι οι μέσοι είναι ίδιοι. Δηλαδή, φαίνεται ότι όντως το χρώμα παίζει ρόλο. Πρέπει στη συνέχεια να εντοπιστύν οι διαφορές. Εφαρμόζοντας τον έλεγχο του Scheffe έχουμε το ακόλουθο αποτέλεσμα:

Multiple Comparisons

Dependent Variable: aritmos_entomon
Scheffe

(I) epifaneia	(J) epifaneia	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	13,50000*	4,10081	,031	,9974	26,0026
	3	5,83333	4,10081	,578	-6,6692	18,3359
	4	4,66667	4,10081	,733	-7,8359	17,1692
2	1	-13,50000*	4,10081	,031	-26,0026	-,9974
	3	-7,66667	4,10081	,348	-20,1692	4,8359
	4	-8,83333	4,10081	,233	-21,3359	3,6692
3	1	-5,83333	4,10081	,578	-18,3359	6,6692
	2	7,66667	4,10081	,348	-4,8359	20,1692
	4	-1,16667	4,10081	,994	-13,6692	11,3359
4	1	-4,66667	4,10081	,733	-17,1692	7,8359
	2	8,83333	4,10081	,233	-3,6692	21,3359
	3	1,16667	4,10081	,994	-11,3359	13,6692

*. The mean difference is significant at the .05 level.

Εικόνα 6: Scheffe Test για την εύρεση μέσων που διαφέρουν

Από τον παραπάνω έλεγχο προκύπτει ότι το πράσινο χρώμα (1) και το κόκκινο (2) διαφέρουν στατιστικά σημαντικά. Επομένως, δεν είναι δυνατόν να θεωρηθεί ότι τα δύο αυτά χρώματα έχουν την ίδια επιρροή.

Το τελευταίο βήμα είναι να καθοριστεί ποια χρώματα μπορούν να κατηγοριοποιηθούν. Το τεστ του Sceffe έδωσε τα εξής αποτελέσματα:

aritmos_entomon

Scheffe[a]

epifaneia	N	Subset for alpha = .05	
		1	2
2	6	36,1667	
3	6	43,8333	43,8333
4	6	45,0000	45,0000
1	6		49,6667
Sig.		,233	,578

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 6,000.

Εικόνα 7: Scheffe Test για ομαδοποίηση κατηγοριών

Παρατηρούμε ότι είτε θα επιλέξουμε τα χρώματα, κόκκινο – κίτρινο - μπλε, είτε τα χρώματα, πράσινο – κίτρινο – μπλε ως ομάδα που τα αποτελέσματα δε διαφέρουν στατιστικά σημαντικά.

Το πιο σημαντικό συμπέρασμα στο οποίο καταλήξαμε είναι ότι όταν είναι όλα τα χρώματα μαζί υπάρχουν στατιστικά σημαντικές διαφορές ως προς τα χρώματα κόκκινο και πράσινο. Διαχωρίζοντας ωστόσο σε δύο ομάδες τα αποτελέσματα δε διαφέρουν σημαντικά.

Για την καλύτερη διεξαγωγή του εκάστοτε πειράματος οφείλουμε να ακολουθούμε κάποιες βασικές αρχές. Έτσι, όταν θέλουμε να ελέγξουμε την επίδραση ενός και μόνο παράγοντα, οφείλουμε να είμαστε σίγουροι ότι οι άλλοι παράγοντες που εμπλέκονται στο πείραμα διατηρούνται σταθεροί. Για το συγκεκριμένο πείραμα άλλοι παράγοντες που ενδεχομένως να επιδρούν είναι:

- Είδη των εντόμων
- Είδος καλλιέργειας
- Χρονική περίοδος διεξαγωγής πειράματος
- Χρονική διάρκεια πειράματος
- Οσμή χρώματος
- Ποσότητα κολλώδους ουσίας
- Γεωγραφική περιοχή

5 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΕΛΕΓΧΟΣ ΑΝΕΞΑΡΤΗΣΙΑΣ

Έχουμε τον πίνακα των δεδομένων:

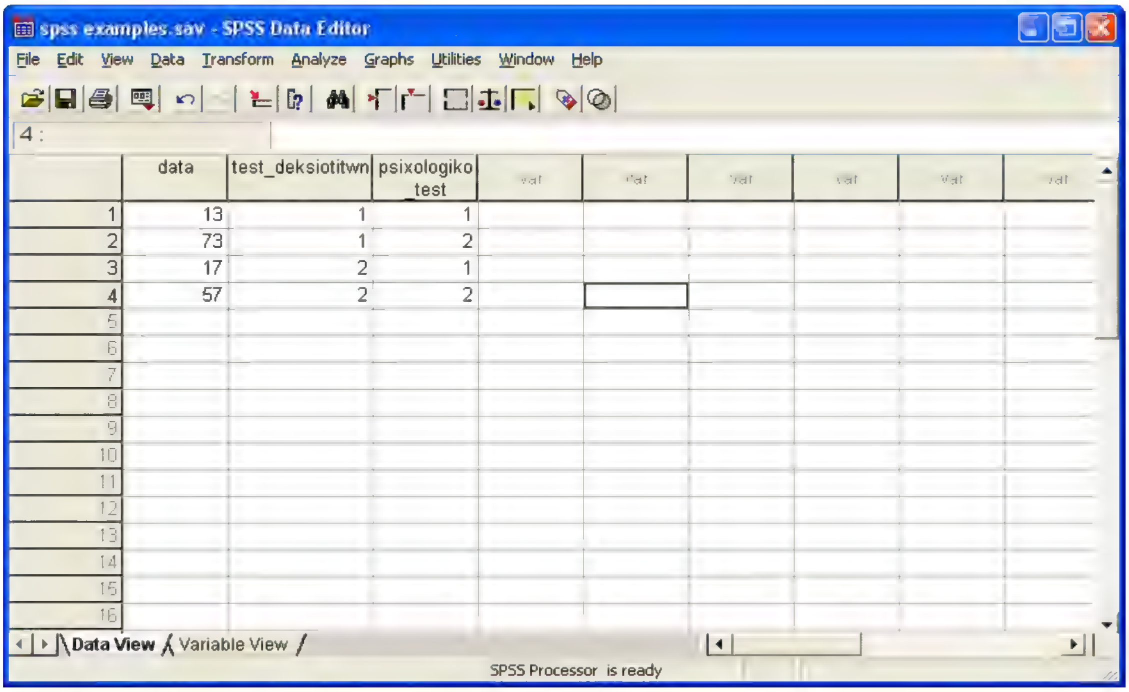
	Ψυχολογικό τεστ		
Τεστ Δεξιοτήτων	**Εσωστρεφείς**	**Εξωστρεφείς**	**Σύνολο**
Επιτυχόντες	13	73	**86**
Αποτυχόντες	17	57	**74**
Σύνολο	**30**	**130**	**160**

Πίνακας 1: Πίνακας δεδομένων

Αρχικά θα πρέπει να καταχωρήσουμε τα δεδομένα μας. Δημιουργούμε τρεις μεταβλητές.

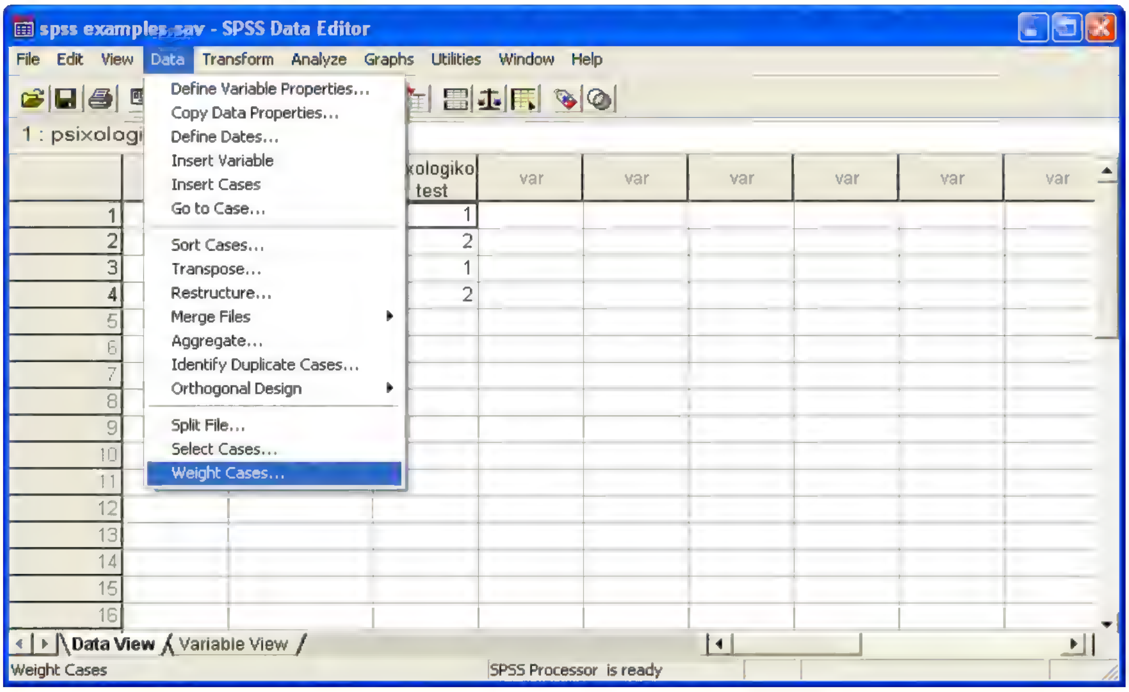
- Η πρώτη θα ονομάζεται ***data*** και θα περιλαμβάνει τα δεδομένα μας, δηλαδή τον αριθμό των ατόμων που ανήκουν σε κάθε κατηγορία.
- Η δεύτερη θα ονομάζεται ***test_deksiotitwn*** και θα παίρνει τις τιμές 1 και 2. 1 αν τα δεδομένα αφορούν τους επιτυχόντες του τεστ δεξιοτήτων και 2 αν αφορούν τους αποτυχόντες.
- Η τρίτη μεταβλητή θα ονομάζεται ***psixologiko_test*** και θα παίρνει τις τιμές 1 εφόσον αφορά 'εσωστρεφείς' πιλότους και 2 εφόσον αφορά 'εξωστρεφείς'.

Τα δεδομένα μας λοιπόν έχουν την παρακάτω μορφή:



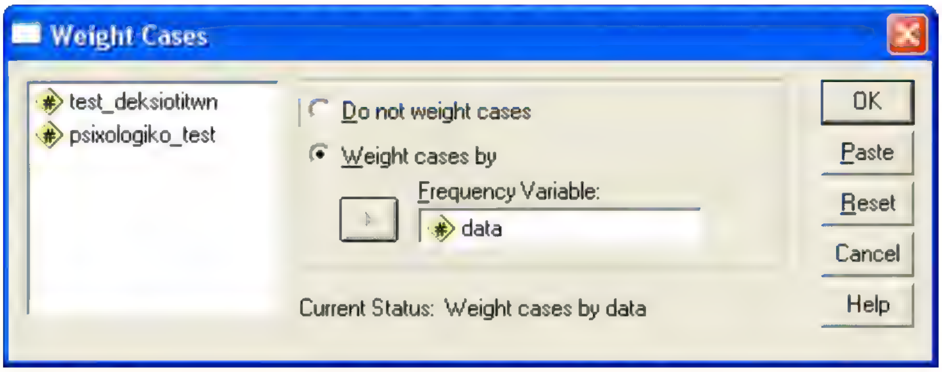
Εικόνα 1 : Εισαγωγή δεδομένων στο SPSS

Είμαστε έτοιμοι λοιπόν να εφαρμόσουμε τη μέθοδο. Ακολουθούμε τα βήματα:



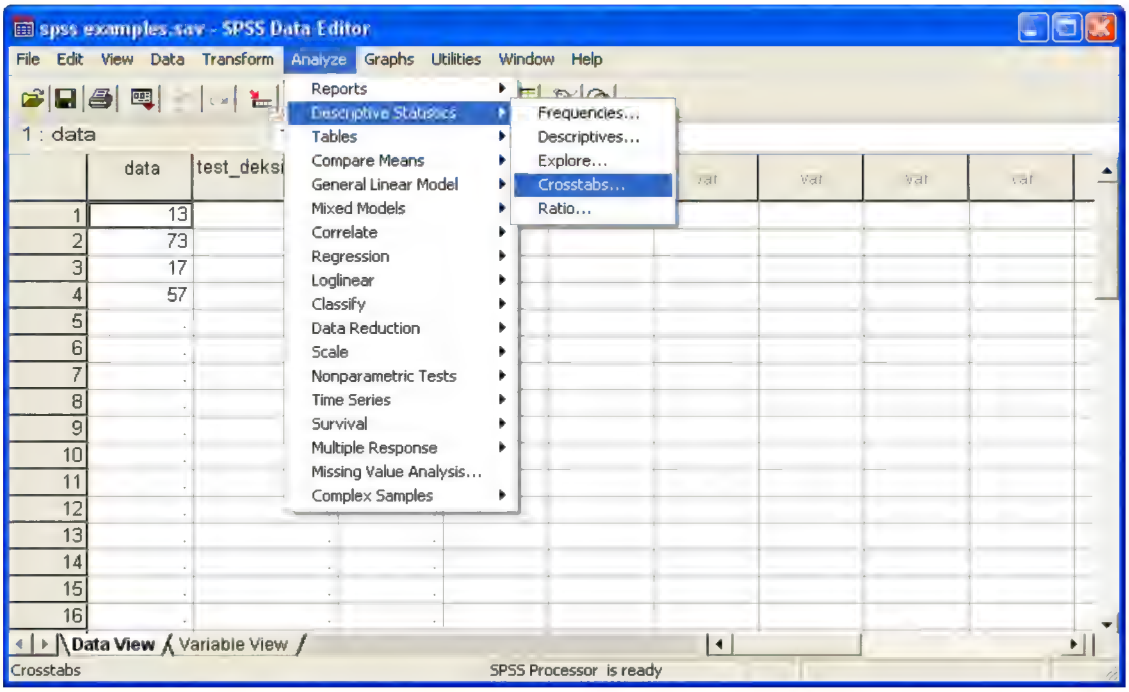
Εικόνα 2 : 1° Βήμα Ελέγχου Ανεξαρτησίας

Σε αυτό το βήμα καθορίζουμε ως προς ποια μεταβλητή σταθμίζουμε τα δεδομένα ώστε να εφαρμόσουμε τον έλεγχο ανεξαρτησίας. Έτσι:



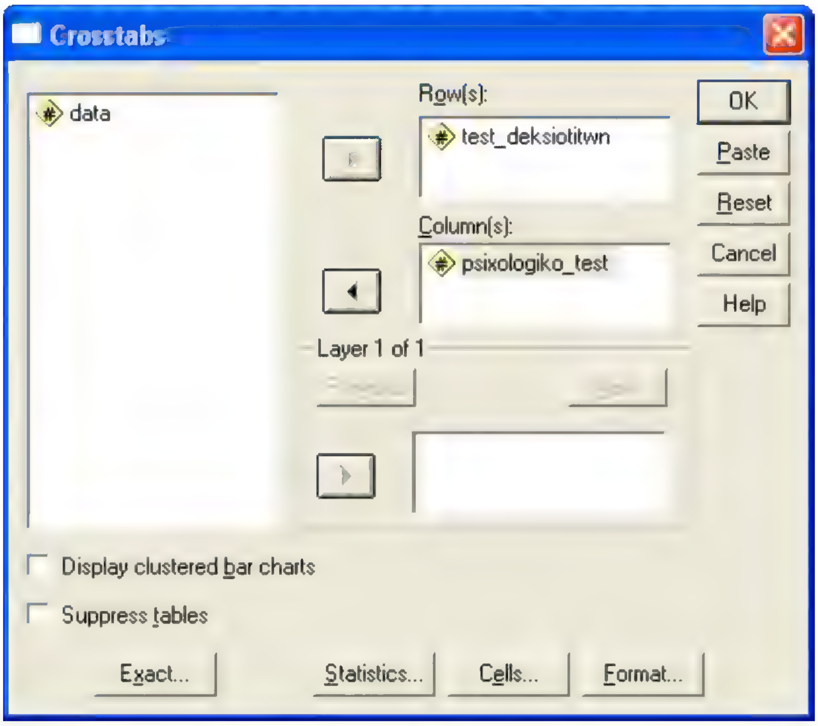
Εικόνα 3: Καρτέλα επιλογής σταθμισμένης μεταβλητής

Θέλουμε ο έλεγχος να γίνει ως προς τον αριθμό των ατόμων που ανήκουν σε κάθε συνδυασμό και για αυτό επιλέγουμε τη μεταβλητή ***data*** ως σταθμισμένη μεταβλητή. Είμαστε έτοιμοι πλέον να εφαρμόσουμε τη διαδικασία. Από το menu επιλέγουμε:



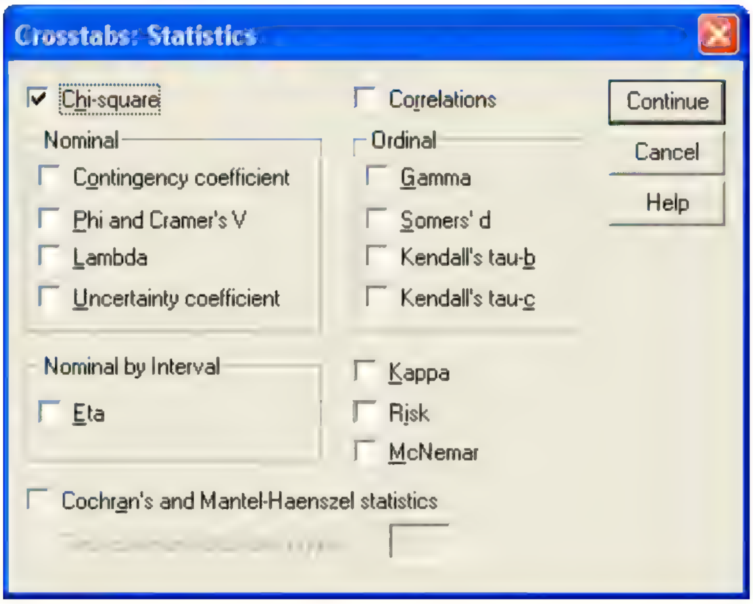
Εικόνα 4: Επιλογή Ελέγχου Συνάφειας

Στο κεντρικό menu καθορίζουμε ποια μεταβλητή θα ανήκει στη γραμμή και ποια στη στήλη του πίνακα που θέλουμε να δημιουργήσουμε. Έτσι:



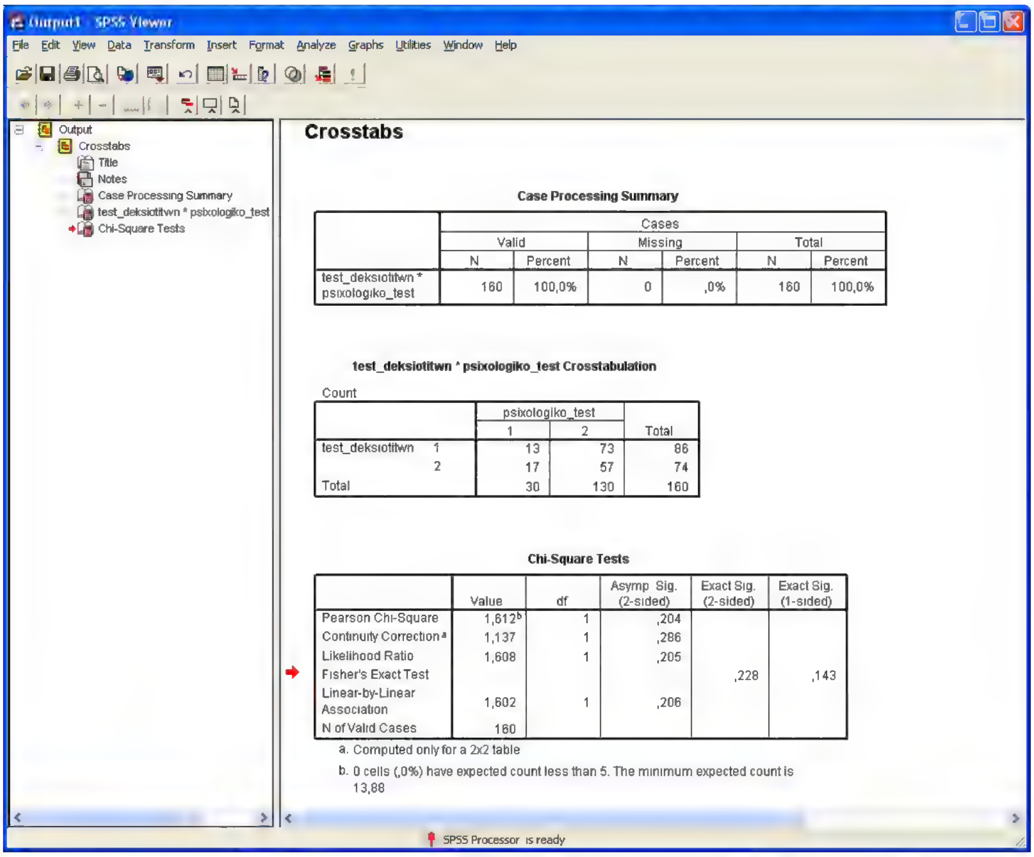
Εικόνα 5: Καρτέλα καθορισμού μεταβλητών

Τέλος, από την καρτέλα ***Statistics*** επιλέγουμε τον έλεγχο ανεξαρτησίας χ^2 για να διαπιστώσουμε τι συμβαίνει μεταξύ των μεταβλητών.



Εικόνα 6: Καρτέλα επιλογής ελέγχου ανεξαρτησίας

Εφαρμόζοντας όλα τα παραπάνω παίρνουμε από το SPSS τα παρακάτω:



Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
test_deksiotitwn * psixologiko_test	160	100,0%	0	,0%	160	100,0%

test_deksiotitwn * psixologiko_test Crosstabulation

Count

		psixologiko_test		Total
		1	2	
test_deksiotitwn	1	13	73	86
	2	17	57	74
Total		30	130	160

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,612[b]	1	,204		
Continuity Correction[a]	1,137	1	,286		
Likelihood Ratio	1,608	1	,205		
Fisher's Exact Test				,228	,143
Linear-by-Linear Association	1,602	1	,206		
N of Valid Cases	160				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 13,88

Εικόνα 7: Πίνακας Συνάφειας 2Χ2

Ο έλεγχος που θέλαμε να κάνουμε είναι:

H_0 : Η ικανότητα του πιλότου είναι ανεξάρτητη από τον τύπο της προσωπικότητας του.

H_1 : Η ικανότητα του πιλότου σχετίζεται με τον τύπο της προσωπικότητας του.

Από τον πίνακα φαίνεται ότι το p-value έχει τιμή 0.204 και αφού αυτή η τιμή είναι μεγαλύτερη από 0.05 (το επίπεδο που πραγματοποιούμε τον έλεγχο) δεν απορρίπτουμε τη μηδενική υπόθεση. Επομένως, δε μπορούμε να θεωρήσουμε ότι η ικανότητα ενός πιλότου σχετίζεται με τον τύπο της προσωπικότητας του.

6 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΣΥΣΧΕΤΙΣΗ

Ερευνητής ενδιαφέρεται να εξετάσει τη σχέση μεταξύ των απαιτούμενων ωρών ύπνου και της ηλικίας των παιδιών. Για το σκοπό αυτό επέλεξε μια ομάδα αποτελούμενη από 13 παιδιά ηλικίας από 4 μέχρι 14 ετών. Στη συνέχεια κατέγραψε την ακριβή ηλικία τους (σε έτη), καθώς και τη μέση χρονική διάρκεια ύπνου σε λεπτά /24ωρο. Ο μέσος χρόνος ύπνου υπολογίστηκε καταγράφοντας για κάθε παιδί το χρόνο ύπνου για τέσσερις διαδοχικές νύχτες υπολογίζοντας στη συνέχεια τη μέση τιμή για τις τέσσερις αυτές τιμές. Τα αποτελέσματα του πειράματος παρουσιάζονται στον πίνακα που ακολουθεί:

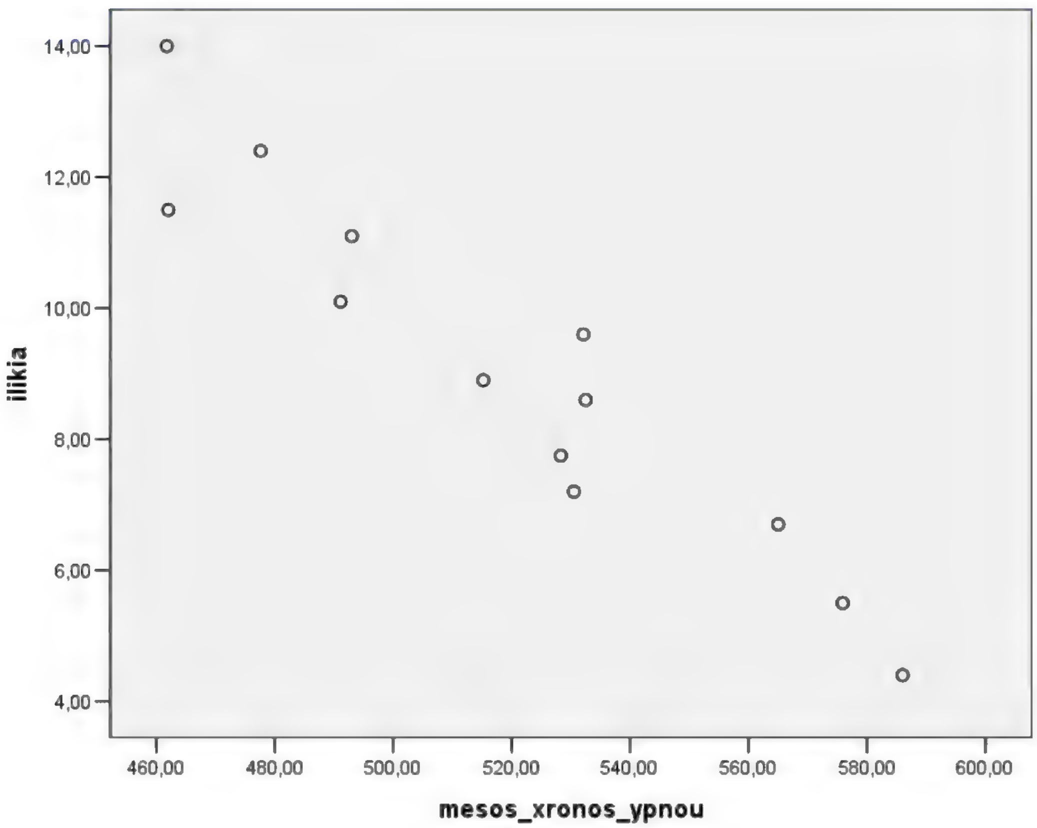
ΗΛΙΚΙΑ	**ΜΕΣΟΣ ΧΡΟΝΟΣ ΥΠΝΟΥ ΣΕ ΛΕΠΤΑ/24ΩΡΟ**
4,40	586,00
14,00	461,75
10,10	491,10
6,70	565,00
11,50	462,00
9,60	532,10
12,40	477,60
8,90	515,20
11,10	493,00
7,75	528,30
5,50	575,90
8,60	532,50
7,20	530,50

Με βάση τα αποτελέσματα του παραπάνω πειράματος να συγγράψετε αναφορά στην οποία θα παρουσιάζετε τα συμπεράσματα που προκύπτουν από αυτά. Η ανάλυση των στοιχείων να γίνει με τη μεθοδολογία που εσείς κρίνετε ως καταλληλότερη, αιτιολογώντας την επιλογή της και ελέγχοντας τις όποιες υποθέσεις απαιτούνται για την εφαρμογή της.

Τέλος, να προτείνετε διάφορους άλλους παράγοντες που ενδεχομένως να επηρεάζουν την χρονική διάρκεια του ύπνου και θα μπορούσαν να ληφθούν υπόψη στο παραπάνω πείραμα.

ΠΡΟΤΕΙΝΟΜΕΝΗ ΑΠΑΝΤΗΣΗ

Η μεθοδολογία που θα χρησιμοποιήσουμε για την επίλυση του προβλήματος αυτού είναι η απλή γραμμική παλινδρόμηση. Εν πρώτης όψεως φαίνεται να υπάρχει εξάρτηση μεταξύ των δύο παραμέτρων. Δηλαδή, είναι λογικό να ισχυριστούμε ότι όσο μεγαλώνει η ηλικία ενός παιδιού τόσο λιγοστεύει και ο απαιτούμενος χρόνος ύπνου. Μέσω του παρακάτω γραφήματος βλέπουμε τι είδος σχέσης έχουν αυτές οι μεταβλητές.



Εικόνα 1: Γράφημα για έλεγχο εξάρτησης

Πράγματι, από το γράφημα φαίνεται να επαληθεύεται μια τέτοιου είδους εξάρτηση. Επομένως, η απλή γραμμική παλινδρόμηση μπορεί να χρησιμεύσει στη συγκεκριμένη περίπτωση. Ως ανεξάρτητη μεταβλητή θα θεωρήσουμε την ηλικία και τον ύπνο ως την εξαρτημένη μας.

Προτού, ωστόσο, προχωρήσουμε στην κατασκευή της θα χρειαστεί να ελέγξουμε κάποιες υποθέσεις για την εφαρμογή της:

- Γραμμικότητα
- Ανεξαρτησία
- Ομοσκεδαστικότητα
- Κανονικότητα

Το μοντέλο μας είναι της μορφής: $\hat{Y} = a + bX$ και ο έλεγχος των παραπάνω υποθέσεων θα γίνει με τη βοήθεια των καταλοίπων.

Από την εφαρμογή του μοντέλου παλινδρόμησης βλέπουμε ότι:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,952[a]	,905	,897	13,15238

a. Predictors: (Constant), ilikia

b. Dependent Variable: mesos_xronos_ypnou

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18220,547	1	18220,547	105,330	,000[a]
	Residual	1902,835	11	172,985		
	Total	20123,382	12			

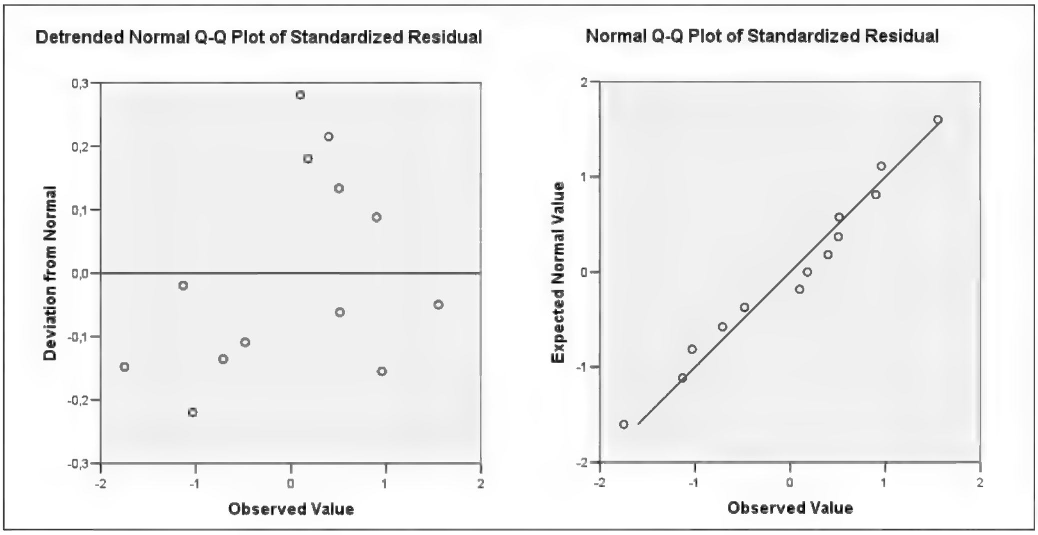
a. Predictors: (Constant), ilikia

b. Dependent Variable: mesos_xronos_ypnou

Εικόνα 2: Έλεγχος Γραμμικότητας

Τόσο το R που είναι υψηλό (0.952), όσο και το p-value που είναι μικρότερο από 0.05 δείχνουν ότι υπάρχει γραμμικότητα στο μοντέλο.

Μέσω των καταλοίπων θα ελέγξουμε τις άλλες υποθέσεις. Έτσι έχουμε:



Εικόνα 3: Έλεγχος ανεξαρτησίας, ομοσκεδαστικότητας και κανονικότητας

Τα κατάλοιπα φαίνεται να κατανέμονται τυχαία και ομοιόμορφα γύρω από το μηδέν, άρα ισχύει τόσο η ανεξαρτησία όσο και η ομοσκεδαστικότητα. Επίσης, φαίνεται ότι η ευθεία γραμμή

προσεγγίζει ικανοποιητικά τα κατάλοιπα και επομένως μπορούμε να πούμε ότι επαληθεύεται και η κανονικότητα.

Εφόσον τελικά ικανοποιούνται όλες οι προϋποθέσεις μπορούμε να προχωρήσουμε στην εξαγωγή του μοντέλου μας. Έχουμε λοιπόν:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	646,483	12,918		50,046	,000	618,052	674,915
	ilikia	-14,041	1,368	-,952	-10,263	,000	-17,052	-11,030

a. Dependent Variable: mesos_xronos_ypnou

Εικόνα 4: Μοντέλο απλής γραμμική παλινδρόμησης

Το μοντέλο μας λοιπόν έχει για a=646.483 και b=-14.041. Φαίνονται και τα διαστήματα εμπιστοσύνης που ενδέχεται να βρίσκεται η πραγματική τιμή του a και του b. Άρα η τελική του μορφή είναι:

Μέσος Χρόνος Ύπνου = 646.48 – 14.04 Ηλικία

Η εξήγηση του μοντέλου είναι η εξής: Αν ηλικία αυξηθεί κατά μία μονάδα τότε ο μέσος χρόνος ύπνου θα μειωθεί κατά 14.04 μονάδες.

Θα πρέπει να σημειώσουμε κάτι πολύ σημαντικό. Το μοντέλο που χρησιμοποιήσαμε της απλής γραμμικής παλινδρόμησης προϋποθέτει την ύπαρξη μια μόνο ανεξάρτητης μεταβλητής. Ωστόσο, ενδεχομένως να υπάρχουν και άλλοι παράγοντες που να επηρεάζουν το μέσο χρόνο ύπνου. Ορισμένοι από αυτούς είναι:

- Ψυχική κατάσταση ατόμου
- Γεύματα
- Απογευματινός ύπνος
- Γεωγραφική περιοχή
- Σωματική κατάσταση
- Χρονική στιγμή που πηγαίνει για ύπνο
- Παθήσεις
- Περιβάλλον

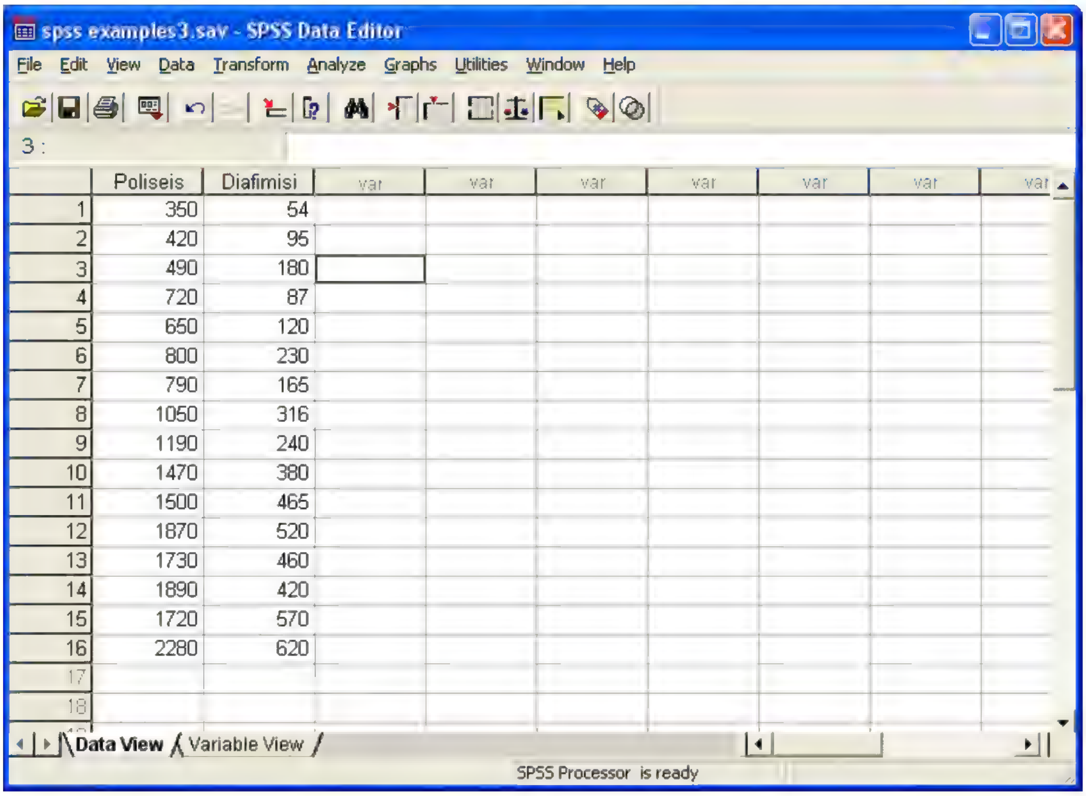
Για ασφαλέστερα συμπεράσματα θα ήταν καλύτερο επίσης να χρησιμοποιείται μεγαλύτερο μέγεθος δείγματος.

7 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ - ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Μία εταιρία θέλει να αυξήσει την κερδοφορία της. Το καλύτερο μοντέλο παλινδρόμησης, που προέκυψε είναι να χρησιμοποιήσουμε τη μεταβλητή 'Έξοδα Διαφήμισης' ως ανεξάρτητη μεταβλητή. Το μοντέλο μας θα είναι της μορφής

'Μέσες Πωλήσεις' = a+b'Έξοδα Διαφήμισης'

Καταχωρούμε τα δεδομένα σε δύο στήλες όπως φαίνεται παρακάτω.

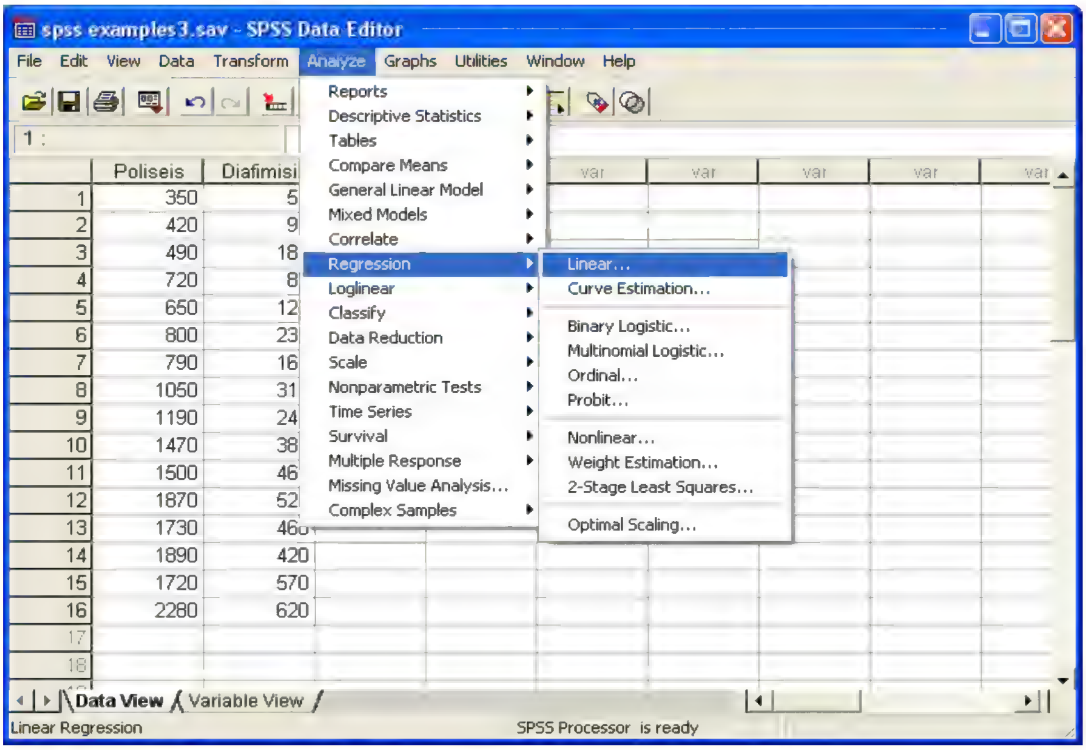


	Poliseis	Diafimisi	var	var	var	var	var	var	var
1	350	54							
2	420	95							
3	490	180							
4	720	87							
5	650	120							
6	800	230							
7	790	165							
8	1050	316							
9	1190	240							
10	1470	380							
11	1500	465							
12	1870	520							
13	1730	460							
14	1890	420							
15	1720	570							
16	2280	620							
17									
18									

Εικόνα 1: Καταχώρηση δεδομένων παραδείγματος

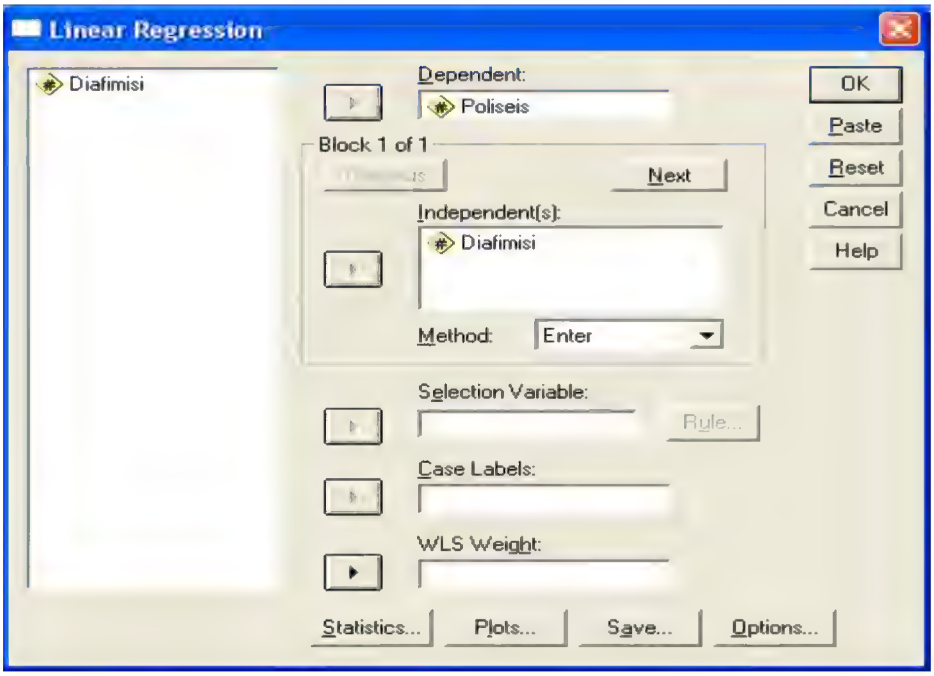
Η μεταβλητή ***Poliseis*** αφορά την εξαρτημένη μας μεταβλητή και η ***Diafimisi*** την ανεξάρτητη. Το επόμενο βήμα είναι να δούμε αν μπορούμε να προχωρήσουμε στην κατασκευή του μοντέλου παλινδρόμησης, αν δηλαδή, ικανοποιούνται οι προϋποθέσεις εφαρμογής του.

Αναλυτικά, ακολουθούμε τα βήματα:



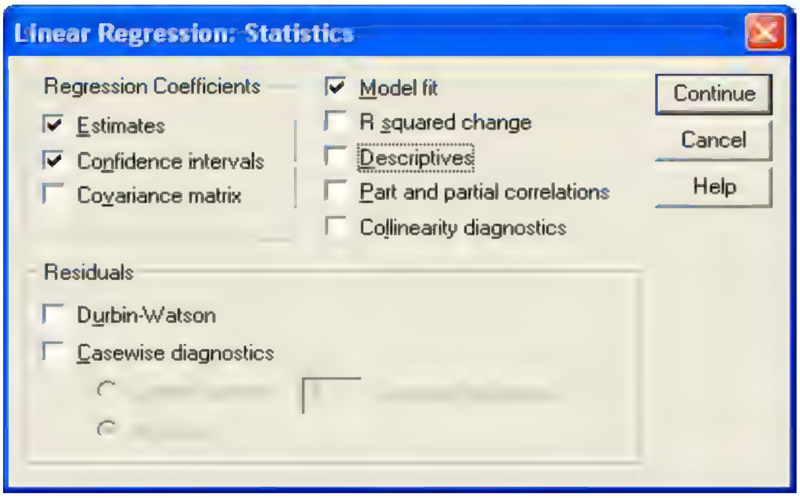
Εικόνα 2: Επιλογή μοντέλου Απλής Γραμμικής Παλινδρόμησης

Επιλέγουμε την εξαρτημένη (***Poliseis***) και την ανεξάρτητη μεταβλητή (***Diafimisi***):



Εικόνα 3: Επιλογή εξαρτημένης και ανεξάρτητης μεταβλητής

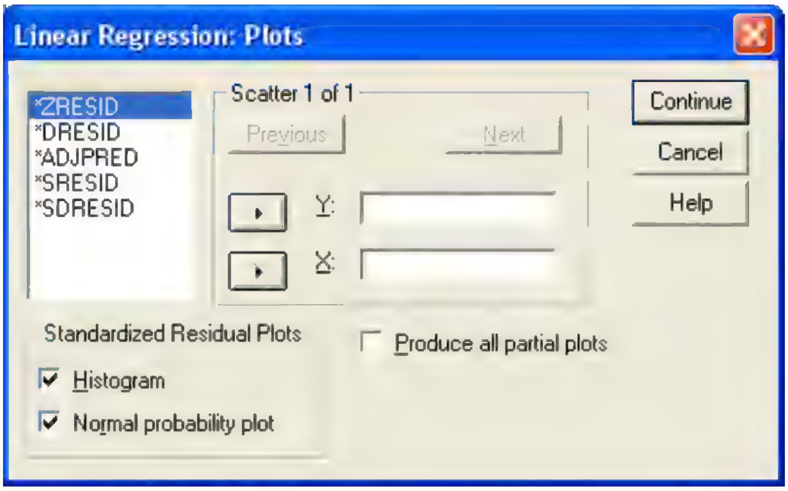
Στη συνέχεια από την καρτέλα ***Statistics*** επιλέγουμε τα εξής:



Εικόνα 4: Καρτέλα *Statistics*

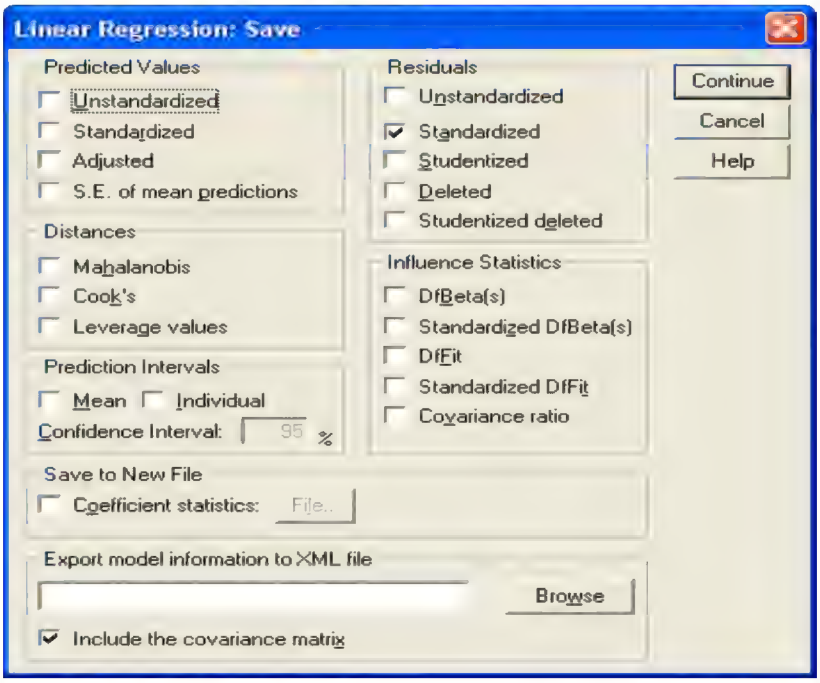
Αν θέλουμε να δούμε για κάθε τιμή του Χ και Υ τα κατάλοιπα αλλά και τις εκτιμηθείσες τιμές του Υ επιλέγουμε και το ***Casewise diagnostics***.

Στη συνέχεια από την καρτέλα ***Plots*** επιλέγουμε τα διαγράμματα των καταλοίπων όπως φαίνεται παρακάτω:



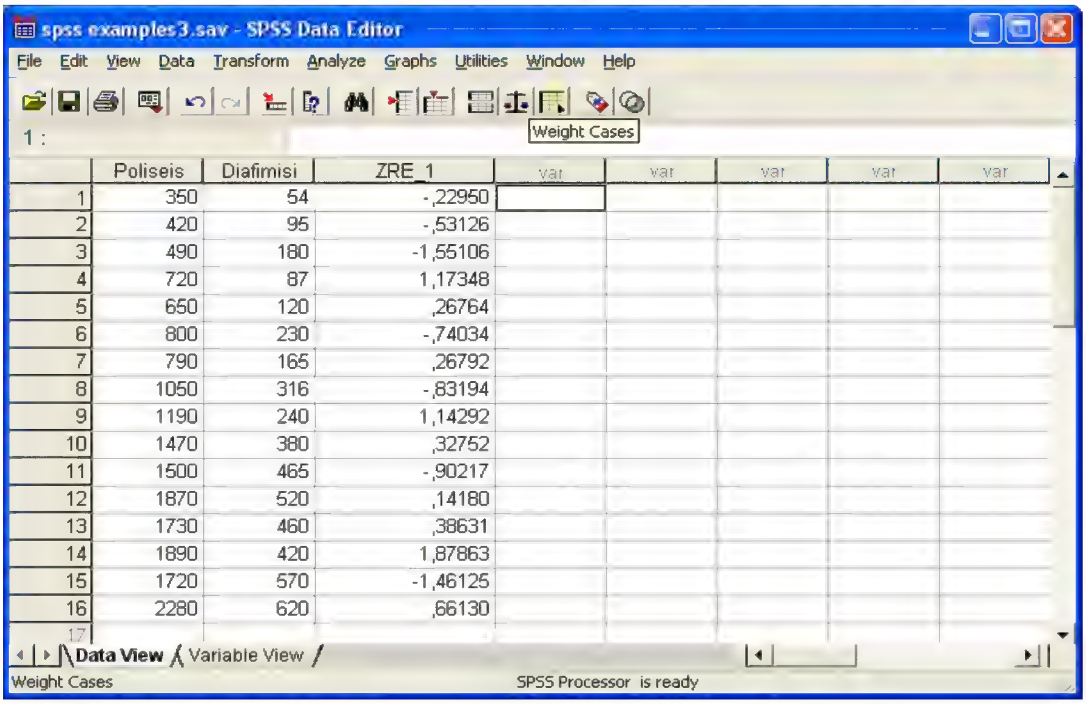
Εικόνα 5: Καρτέλα *Plots*

Στην καρτέλα ***Save*** καθορίζουμε τι θέλουμε να αποθηκευτεί στο κεντρικό menu των μεταβλητών μας. Επιλέγουμε τα ***standardized residuals***.



Εικόνα 6: Καρτέλα *Save*

Πλέον, στο κεντρικό menu έχει προστεθεί μία ακόμη μεταβλητή που αφορά τα κατάλοιπα:



Εικόνα 7 : Κεντρικό menu μεταβλητών

Όπως είπαμε θα χρησιμοποιήσουμε τα κατάλοιπα για να δούμε αν μπορούμε να προχωρήσουμε με το μοντέλο της παλινδρόμησης. Γενικά για την εφαρμογή ενός μοντέλου παλινδρόμησης πρέπει να ισχύουν τα εξής για τα κατάλοιπα:

- Ανεξαρτησία
- Ομοσκεδαστικότητα
- Κανονικότητα

Για τον έλεγχο της ανεξαρτησίας θα χρησιμοποιήσουμε το Runs Test. Το αποτέλεσμα που παίρνουμε είναι το ακόλουθο:

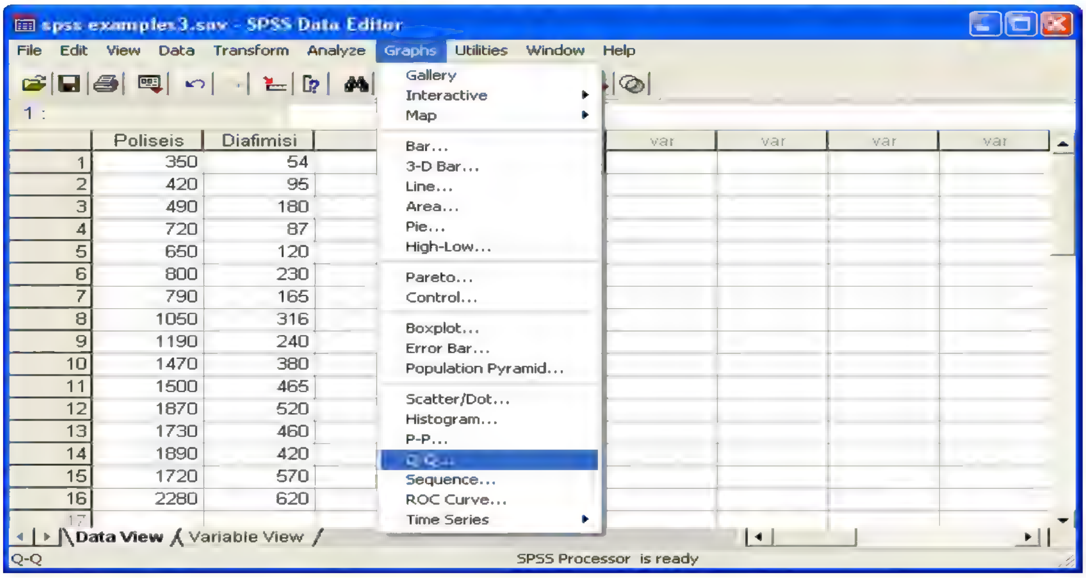
Runs Test

	Standardized Residual
Test Value [a]	,20472
Cases < Test Value	8
Cases >= Test Value	8
Total Cases	16
Number of Runs	10
Z	,259
Asymp. Sig. (2-tailed)	,796
Exact Sig. (2-tailed)	,810
Point Probability	,190

a. Median

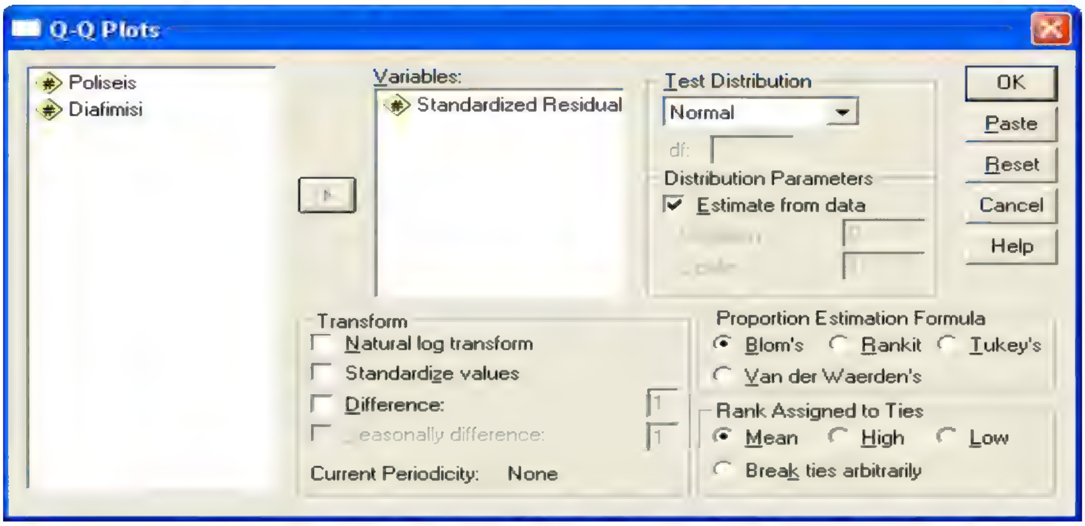
Εικόνα 8: Runs Test για έλεγχο ανεξαρτησίας καταλοίπων

Το p-value έχει τιμή 0.796, μεγαλύτερη από 0.05 και επομένως δεν απορρίπτουμε τη μηδενική υπόθεση περί ανεξαρτησίας – τυχαιότητας των καταλοίπων. Για τις υπόλοιπες υποθέσεις έχουμε:

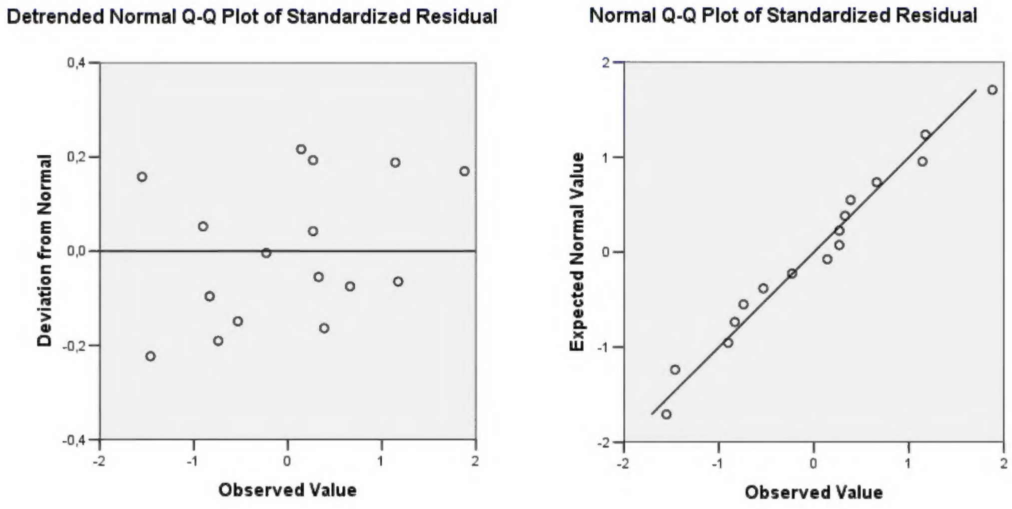


Εικόνα 9: Επιλογή ελέγχου καταλοίπων

Επιλέγουμε τη μεταβλητή μας και παίρνουμε το εξής γράφημα



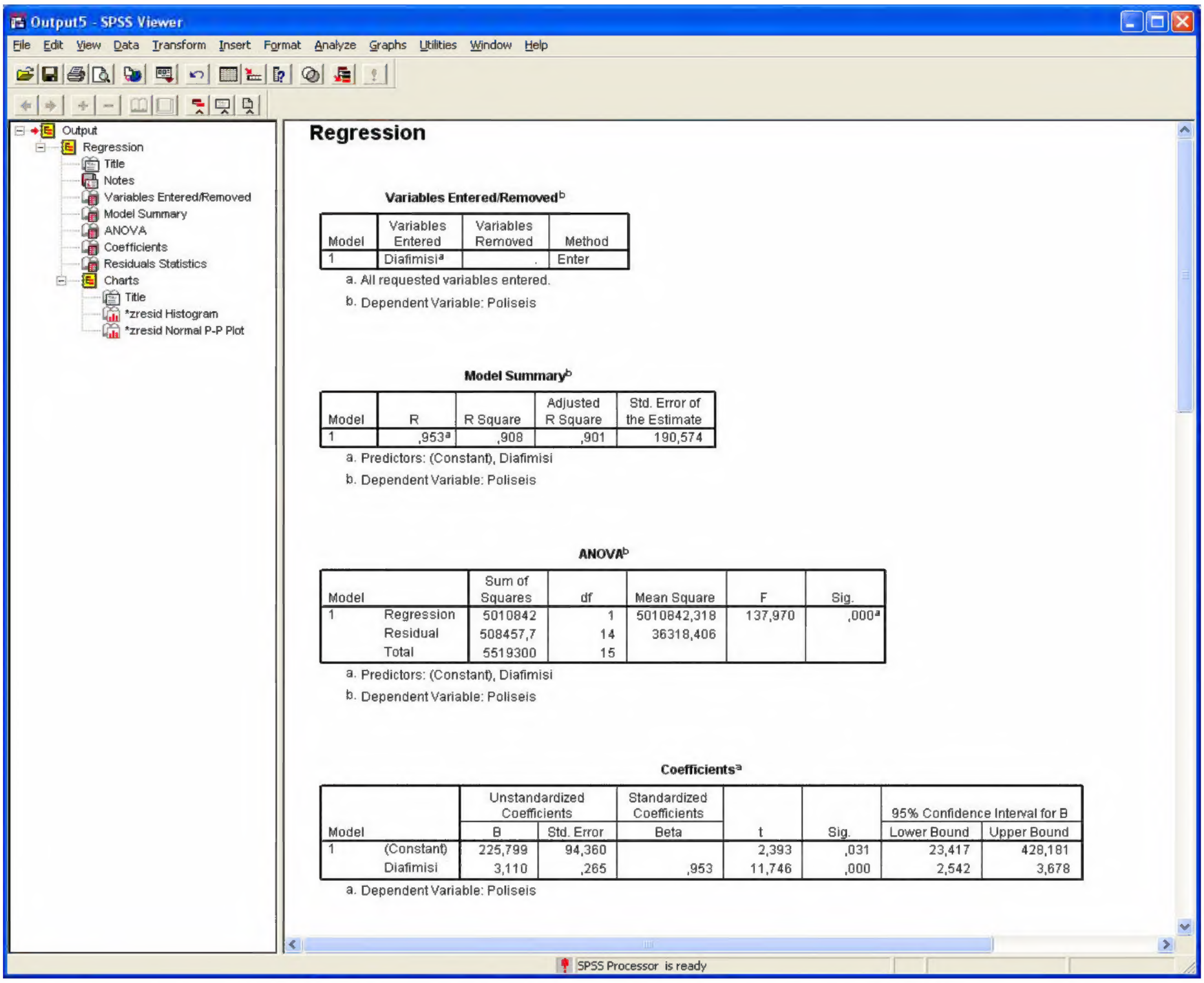
Εικόνα 10: Επιλογή καταλοίπων για κατασκευή γραφημάτων



Εικόνα 11: Διαγράμματα καταλοίπων για έλεγχο κανονικότητας και ομοσκεδαστικότητας

Από τα πρώτο γράφημα φαίνεται ότι τα κατάλοιπα κατανέμονται τυχαία γύρω από το μηδέν και επομένως φαίνεται να ισχύει η ομοσκεδαστικότητα. Επίσης, στο δεύτερο γράφημα φαίνεται ότι τα κατάλοιπα δεν απέχουν πολύ από τη γραμμή που δείχνει την κανονικότητα και επομένως μπορούμε να πούμε ότι ισχύει και αυτή η υπόθεση.

Αφού λοιπόν ισχύουν οι παραπάνω υποθέσεις μπορούμε να προχωρήσουμε ανάλυση του μοντέλου που προέκυψε.



Εικόνα 12: Μοντέλο Απλής Γραμμικής Παλινδρόμησης

Αναλυτικά:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,953^a	,908	,901	190,574

a. Predictors: (Constant), Diafimisi

b. Dependent Variable: Poliseis

Εικόνα 13 : Έλεγχος γραμμικότητας μοντέλου

Βλέπουμε ότι ο συντελεστής συσχέτισης είναι 0.953, πολύ υψηλός. Επίσης, τόσο το r^2 (συντελεστής προσδιορισμού) όσο και το r^2_{adj} έχουν τιμή που τείνει στο 1 επομένως έχουμε ενδείξεις για τη γραμμικότητα του μοντέλου. Μάλιστα από τον πίνακα της ANOVA βλέπουμε τον

έλεγχο που αφορά αν ο συντελεστής προσδιορισμού είναι στατιστικά σημαντικός. Το p-value είναι πρακτικά ίσο με μηδέν και επομένως απορρίπτουμε τη μηδενική υπόθεση ότι το $r^2 = 0$.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	225,799	94,360		2,393	,031	23,417	428,181
	Diafimisi	3,110	,265	,953	11,746	,000	2,542	3,678

a. Dependent Variable: Poliseis

Εικόνα 14 : Τιμές στις παραμέτρους του μοντέλου

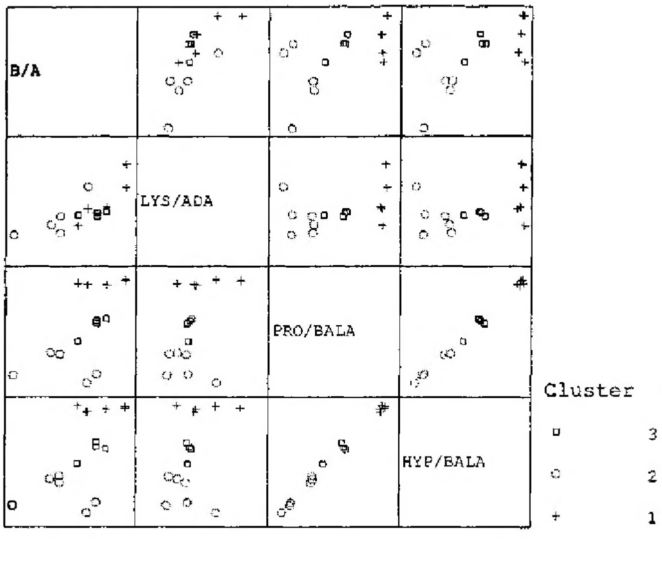
Παρατηρούμε ότι το a έχει τιμή 225,799 ενώ το b 3,110. Άρα το μοντέλο παλινδρόμησης είναι το:

‘Μέσες Πωλήσεις’ = 225.799+3,11‛Έξοδα Διαφήμισης’

Επίσης, πραγματοποιείται και ο έλεγχος σημαντικότητας για την κάθε μία παράμετρο του μοντέλου. Και για τις δύο παραμέτρους τα p-values είναι μικρότερα από 0.05 (0.031 και ≈0 αντίστοιχα) και επομένως απορρίπτουμε τις μηδενικές υποθέσεις ότι οι τιμές των παραμέτρων a, b είναι μηδέν.

Τέλος, δίνονται και διαστήματα εμπιστοσύνης για την κάθε μία παράμετρο. Δηλαδή, ένα 95% διάστημα εμπιστοσύνης για το a είναι (23.417 , 428.181) ενώ για το b το αντίστοιχο διάστημα εμπιστοσύνης είναι το (2.542 , 3.678).

Πίνακας (1) : Απεικονίζει τα αποτελέσματα της K-means ομαδοποίησης



Εικόνα (1):

Πολλαπλό Διάγραμμα σημείων για τις μεταβλητές που χρησιμοποιήσαμε

Βλέπουμε λοιπόν πως η πρώτη ομάδα (+) 1 έχουν μεγάλες τιμές για τη μεταβλητή HYP/BALA και PRO/BALA και B/A. Η ομάδα 3 (απεικονίζεται με ένα τετραγωνάκι) είναι για όλες τις μεταβλητές σε μια μεσαία κατάσταση ενώ η ομάδα 2 (απεικονίζεται με ένα κυκλάκι) έχει σε όλες τις μεταβλητές τις μικρότερες τιμές.